Ex-post legislative evaluations in the European Commission

van Voorst, Stijn

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Ex-post legislative evaluations in the European Commission

Between technical instruments and political tools

"Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op woensdag 19 december 2018 om 16.00 uur door Stijn van Voorst, geboren te Tilburg".

Promotores: Prof. Dr. A.C.M. Meuwese

Prof. Dr. E. Mastenbroek

Prof. dr. S. van Thiel

Promotiecommissie: Prof. Dr. M. L. P. Groenleer

Dr. Ir. T. Havinga

Prof. Dr. V. Mak

Prof. Dr. C. M. Radaelli

Prof. Mr. L. A. J. Senden

Prof. Dr. B. Steunenberg

Table of Contents

Preface

Six years ago, in the spring of 2012, I was first introduced to the topic of ex-post evaluations of EU legislation. At first sight the topic did not strike me as particularly engaging: I was simply interested in European policies and the only project related to that which my master programme offered concerned evaluations. Only when I delved into the subject I found out that there was much more to these evaluations than just technical exercises: at the heart of the matter were all sorts of interesting questions regarding the political interests of European policy-makers, their accountability towards citizens and the use of objective information to improve how EU legislation affects citizens and companies.

In the six subsequent years I spent the majority of my working time on studying ex-post evaluations of EU legislation from an academic perspective, first as a master student, then as a junior researcher and finally as a PhD student from September 2014 onwards. The result of all this work is this dissertation, in which I attempt to provide a comprehensive overview of and explanation for the variation among the evaluations.

There are many people who deserve credit for helping me to complete this dissertation. The most important of those are my three supervisors: prof. dr. Ellen Mastenbroek, prof. dr. Anne Meuwese and prof. dr. Sandra van Thiel. Their in-depth feedback and continuous support greatly contributed to the quality of my work. I also wish to sincerely thank Thomas van Golen LLM MSc and dr. Pieter Zwaan for all their help. Not only are they co-authors of various articles included in this dissertation, but their comments also helped me to improve other parts of the PhD thesis.

Furthermore, I would like to thank the many respondents within the European Commission and other organizations that provided data that I could use for my research. Throughout my PhD process I have been continuously impressed by the openness of the EU institutions to my questions and requests. In particular I would like to mention Daniel Klein of the Commission's Directorate-General for Competition, who provided extensive information

about the situation in his own organization and took the time to give useful advice about various parts of my theoretical framework.

Others made smaller or less direct contributions, yet in the end their help was just as crucial to complete this dissertation. Prof. Dr. Claudio Radaelli was kind enough to receive me at the University of Exeter for a month during my PhD process, which allowed me to discuss many questions regarding better regulation and evaluation use with him and other academics at his department. Dr. Peter Kruyen provided answers to several detailed statistical questions that I had regarding the quantitative part of my research. Sebastian Lemire and Thomas Delahais helped me to measure the abstract concept of evaluation capacity, which contributed to many chapters of this dissertation. Korné Boerman provided useful feedback on various parts of my writing. I would like to thank all these people for their generous and useful assistance.

Stijn van Voorst
June 2018

Summary

Introduction of the topic

Since the year 2000, the European Commission has repeatedly formulated the ambition to systematically evaluate all major EU legislation. In 2003 this ambition resulted in the introduction of *impact assessments*: reports assessing the costs and benefits of legislative proposals. From 2007 onwards the Commission also started to systematically conduct ex-post legislative (EPL) evaluations: reports assessing the functioning of regulations and directives currently in force. Some EPL evaluations only study the transposition of EU directives to national legislation or their practical implementation; other reports (also) assess the intended and unintended effects of EU legislation on society.

Together with impact assessments and public consultations, EPL evaluations are the main components of the Commission's Better Regulation Agenda. In theory such evaluations may fulfil two important functions related to EU legislation. Firstly, by recommending how the implementation of legislation can be improved and/or how legislation can be amended to increase its effectiveness, EPL evaluations are a potential tool for decision-makers to improve their policies. Secondly, EPL evaluations can be used by actors like the European Parliament (EP) and the Council of Ministers to hold the Commission accountable for its decisions related to legislative implementation. For example, these actors can ask the Commission critical questions based on evaluation results.

Results per topic

This dissertation presents the first large-scale academic research about the Commission's EPL evaluations. Its key assumption is that such evaluations can only contribute to learning and accountability if they meet three conditions: systematic initiation, high quality and systematic use. The main goal of this dissertation is therefore to describe and explain the variation in the initiation, quality and use of the Commission's EPL evaluations, to assess to what extent the Commission's system for such evaluations is fit to enhance learning and accountability.

The first condition, **systematic initiation**, means that all major legislation should be evaluated periodically. Although EPL evaluations may lead to the improvement of specific legislation even if this requirement is not met, in that case they will not enhance legislative quality as a whole. If the Commission conducts EPL evaluations selectively it could also create the impression that it decides what legislation to evaluate based on political motives. Such a reputation could harm the credibility of all its subsequent evaluations.

Chapter 4 of this dissertation shows that about 42% of all major EU legislation from 2000-2004 has been evaluated ex-post by the Commission. This means that more than half of the major EU legislation from 2000-2004 has never been evaluated. These findings reveal that the Commission only partly meets the requirement of systematic initiation.

Four factors significantly affect the variance in the initiation of EPL evaluations by the Commission. First, the *type of legislation* matters: directives are more likely to be evaluated than regulations. Second, the chances that a piece of legislation is evaluated increase with its *complexity*. Both of these explanations suggest that the Commission may prioritize evaluating legislation that grants more freedom to the member states, because for such legislation the risk of non-compliance is higher. In other words, EPL evaluations may partly be initiated by the Commission to make its task of enforcing EU legislation easier.

A third significant explanation for the variance in the initiation of EPL evaluations by the Commission is the *presence of evaluation clauses*: Legislation containing a provision that requires it to be evaluated within a given number of years is much more likely to be evaluated than legislation without such a provision. The fourth significant explanation for the variance in the initiation of EPL evaluations is the *evaluation capacity* of the responsible Directorate-General (DG). DGs are the main organizational components of the Commission and have considerable freedom in their evaluation policies. DGs with a specialized unit for ex-post evaluations and/or specific guidelines for EPL evaluations turned out to evaluate a significantly higher proportion of their legislation than other DGs.

The second condition, **high quality**, means that EPL evaluations can only contribute to learning and accountability if they meet certain methodological standards. If the Commission's EPL evaluations are not valid and reliable, any decisions that take these evaluations into account are based on misleading information. Furthermore, a lack of quality can create the

perception among decision-makers that evaluation findings misrepresent reality, which makes it less likely that such findings will be used for learning in the future.

Chapter 5 of this dissertation shows that the quality of the Commission's EPL evaluations that assess effectiveness varies considerably. The vast majority (76%) of the reports that were studied used a robust combination of stakeholder input and other forms of data collection. However, the evaluations perform less well regarding other aspects of quality. Whereas almost all reports (89%) have a well-defined scope in the sense of clearly specified research questions, less than 40% of them go beyond this by also describing the intervention logic of the legislation that they evaluate. Between 40% and 70% of the EPL evaluations meet criteria like the presence of a clear operationalization (internal validity), a clear country selection and a clear case selection (external validity) and the presence of substantiated conclusions. By far the worst aspect of the evaluations' quality is their replicability: only 31% of the reports contained or referred to all the material that would be required to repeat the underlying research, like interview guides and lists of respondents.

The key determinant for this variance in evaluation quality is the *type of evaluator*: EPL evaluations conducted by external consultants are of significantly higher quality than evaluations conducted internally by the Commission. This suggests that the technical expertise of external parties is a crucial asset when it comes to properly evaluating EU legislation. The evaluation capacity of the Commission's DGs, the complexity of the evaluated legislation and various political conditions were found to have no effect on the variance in quality. The results do show that evaluations of legislation that had to be approved by the Europees Parlement (EP) are of higher quality than other evaluations, but more research is needed to find out why that causal relation exists.

The third condition, **systematic use**, means that the results of EPL evaluations need to be seriously considered during decision-making moments. If this requirement is not met, the evaluations are essentially a waste of time and money, as without use there is no way in which they can contribute to learning and accountability.

Chapter 6 of this dissertation shows that the results of the Commission EPL's evaluations are frequently used in impact assessments (evaluations of the costs and benefits of legislative proposals). About 65% of the impact assessments for which a prior EPL evaluation is available

make use of that evaluation, although the level of use varies from making a single reference to an in-depth forms of analysis. The *timeliness* of the EPL evaluations turns out to be a necessary condition for their use in impact assessments.

Chapter 7 of this dissertation studies the effect of political conditions on the use of the Commission's EPL evaluations for the purpose of learning. The results falsify the hypothesis that such use varies based on the preferences of actors that the Commission depends on, such as the European Parliament, the Council and major interest groups. Instead, it turns out that the *Commission's own political priorities* are the most important explanation for use. Ever since the Juncker Commission entered into office in 2014, the institution has become more reluctant to propose new legislation, in part as a response to criticism by Eurosceptics. Especially in policy fields that are no priority of the current Commission it has therefore become difficult to translate the results of EPL evaluations in policy changes. Conversely, in policy fields that are political priorities of the current Commission, there is much opportunity for EPL evaluations to contribute to learning.

Chapter 8 of this dissertation addresses the use of the Commission's EPL evaluations in questions of the European Parliament. In theory, evaluations are a useful source of information for parliamentarians to hold the Commission accountable for its decisions. However, in practice only 22% of the EPL evaluations studied in this dissertation turned out to be mentioned in any EP questions. The only significant explanation for variation in this regard is the *level of conflict between the EP and the Commission*: the chances that an evaluation is used in questions of the EP is significantly higher for evaluations of topics that were controversial during the legislative process than for evaluations of other topics.

To place the results presented above into perspective, it should be noted that most OECD countries do not have systematic procedures for EPL evaluations at all, which means that the Commission outperforms them by default. Furthermore, even the few OECD countries that have systematic procedures for EPL evaluations in place appear to face problems concerning their initiation, quality and use, which shows that such issues are not unique to the Commission. Therefore, the Commission is clearly ahead of or on par with most national systems for EPL evaluations.

General conclusions

Various academic literature suggests that the European Commission is (partly) driven by its interest to maximize its competences. When applied to EPL evaluations, this theory leads to the hypothesis that the initiation and quality of such evaluations are lower in those cases where the Commission perceives a higher risk that negative evaluation results could lead to criticism on its competences. However, the results of this dissertation do not confirm this hypothesis. They do show that various other political and technical factors affect the initiation, quality and use of the Commission's EPL evaluations. These factors vary considerably from subject to subject and have therefore already been summarized above.

Besides these theoretical implications, the results of this dissertation have some practical implications as well. First, the findings show that evaluation clauses can be a useful tool to encourage the systematic initiation of EPL evaluations in the EU (although they appear to have no effect on evaluation quality). Second, the results reveal that extra investments in evaluation capacity can help the Commission to evaluate a larger proportion of EU legislation. Third, the results show that the timely availability of EPL evaluations is crucial to allow their results to be used in impact assessments, which shows the importance of strictly enforcing the Commission's 'evaluate first' principle.

In conclusion, whereas the Commission's current system for EPL evaluations contributes to learning and accountability to some extent, significant further developments regarding the initiation, quality and use of these evaluations appear to be necessary for these benefits to become more systematic. Hopefully, the specific findings and recommendations presented in this dissertation can contribute to such improvements. In this day and age when EU legislation increasingly affects that day-to-day activities of citizens and companies and is frequently criticized by Eurosceptic actors, it is all the more important to ensure a continuous stream of reliable information about the functioning of such legislation is available. If EPL evaluations can fulfil this role, they may contribute to step-by-step improvements to the effects of legislation, the democratic accountability of the EU's institutions, and the legitimacy of the European project as a whole.

Samenvatting

Introductie van het onderwerp

Sinds het jaar 2000 heeft de Europese Commissie herhaaldelijk de ambitie uitgesproken om alle belangrijke wetgeving van de Europese Unie (EU) systematisch te evalueren. In 2003 resulteerde deze ambitie in het opzetten van een systeem voor zogenaamde *impact assessments*: rapporten die de verwachte kosten en baten van wetgevingsvoorstellen beoordelen. Vanaf 2007 begon de Commissie ook met het systematisch uitvoeren van ex-post wetgevingsevaluaties (vanaf nu: EPL evaluaties): rapporten die reeds bestaande Europese verordeningen en richtlijnen beoordelen. Sommige EPL evaluaties beoordelen slechts de omzetting van Europese richtlijnen naar nationale wetgeving of hun implementatie in de praktijk; andere evaluaties bestuderen (ook) de gewenste en ongewenste maatschappelijke effecten van de wetgeving.

Samen met impact assessments en openbare consultaties vormen EPL evaluaties de belangrijkste bouwstenen van de Agenda voor Betere Regelgeving van de Commissie. In theorie vervullen zulke evaluaties namelijk minstens twee belangrijke functies rond het Europese wetgevingsproces. Ten eerste is dit de functie van *leren*: de rapporten leveren informatie op over de implementatie, naleving en maatschappelijke effecten van Europese regels, die de Europese Commissie vervolgens kan gebruiken als basis voor besluitvorming over de verbetering van deze wetgeving. Ten tweede spelen EPL evaluaties een rol bij het afleggen van (democratische) verantwoording: via hun resultaten kunnen actoren zoals het Europees Parlement en de Raad van Ministers de acties van de Europese Commissie rond de uitvoering van wetgeving te beoordelen. Op basis van hun oordeel kunnen deze actoren vervolgens proberen het gedrag van de Commissie bij te sturen, bijvoorbeeld door het stellen van kritische vragen naar aanleiding van evaluatieresultaten.

Resultaten per deelonderwerp

Deze dissertatie presenteert het eerste grootschalige academisch onderzoek dat heeft plaatsgevonden naar de EPL evaluaties van de Europese Commissie. De centrale assumptie van

het onderzoek is dat zulke evaluaties alleen kunnen bijdragen aan leren en verantwoording als ze voldoen aan drie voorwaarden: systematische initiëring, hoge kwaliteit en systematisch gebruik. Het hoofddoel van deze dissertatie is dan ook het beschrijven en verklaren van de variantie in de initiëring, de kwaliteit en het gebruik van de EPL evaluaties van de Commissie, om zo te kunnen beoordelen in hoeverre en waarom het systeem van de Commissie al dan niet aan de gestelde voorwaarden voldoet.

De eerste voorwaarde, **systematische initiëring**, betekent dat alle belangrijke wetgeving periodiek moet worden geëvalueerd. Als deze voorwaarde wordt geschonden leiden EPL evaluaties wellicht tot leren en verantwoording voor een beperkt deel van de Europese wetgeving, maar vinden deze baten niet plaats over de gehele linie. Een gebrek aan systematische initiëring kan bovendien de verdenking scheppen dat de Commissie selectief evaluaties uitvoert op basis van de verwachte resultaten, wat de geloofwaardigheid van het hele systeem voor EPL evaluaties onderuit kan halen.

De resultaten van hoofdstuk 4 van deze dissertatie tonen aan dat circa 42% van de belangrijke EU wetgeving uit de jaren 2000-2004 is geëvalueerd door de Commissie. Dit betekent dat meer dan de helft van de belangrijke Europese wetgeving uit die jaren niet is geëvalueerd en dat de Commissie dus slechts ten dele voldoet aan de voorwaarde van systematische initiëring. Wel lijkt de proportie belangrijke wetgeving die de Commissie evalueert met de tijd toe te nemen.

Vier factoren blijken te verklaren waarom de Commissie sommige wetgeving wel evalueert en andere wetgeving niet. Ten eerste is dit het *type wetgeving*: richtlijnen hebben een veel grotere kans te worden geëvalueerd dan verordeningen. Ten tweede is de *complexiteit van de wetgeving* een verklaring: hoe ingewikkelder de wetgeving, hoe groter de kans op een evaluatie. Deze resultaten suggereren dat de Commissie mogelijk prioriteit geeft aan het evalueren van wetgeving waarbij de kans op gebrekkige naleving door de lidstaten van de EU groter is. Zowel bij richtlijnen als bij relatief complexe wetgeving hebben nationale overheden namelijk doorgaans veel ruimte om de uitvoering zelf vorm te geven. EPL evaluaties kunnen in zulke situaties een nuttige bron van informatie zijn voor de Commissie om te achterhalen welke landen de wetgeving (niet) naleven.

Een derde factor die de variatie in de initiëring van EPL evaluaties verklaart is de *aanwezigheid van evaluatieclausules*: artikelen in EU wetgeving die een evaluatie na een bepaald aantal jaren verplichten. De Commissie blijkt wetgeving met een dergelijke clausule veel vaker te evalueren dan andere wetgeving, hoewel er ook veel wetgeving met een clausule bestaat die niet wordt geëvalueerd. De vierde verklarende factor is de *evaluatiecapaciteit van de betrokken directoraten-generaal (DGs)*. DGs zijn de belangrijkste organisatorische componenten van de Commissie; in de praktijk hebben zij veel vrijheid bij het vormgeven van hun eigen evaluatiebeleid. De resultaten van deze dissertatie laten zien dat DGs die meer middelen in EPL evaluaties stoppen en betere procedures voor zulke evaluaties hebben een groter percentage van hun wetgeving evalueren.

De tweede voorwaarde, **hoge kwaliteit**, houdt in dat EPL evaluaties alleen kunnen bijdragen aan leren en verantwoording als ze voldoen aan standaarden voor degelijk onderzoek. Als niet aan deze voorwaarde wordt voldaan kloppen de conclusies van EPL evaluaties waarschijnlijk niet, waardoor eventuele besluiten die naar aanleiding van de evaluaties worden genomen op verkeerde informatie zijn gebaseerd. Ook kan bij gebrekkige kwaliteit de geloofwaardigheid van alle toekomstige EPL evaluaties verloren gaan.

Hoofdstuk 5 van deze dissertatie toont aan dat de kwaliteit van de EPL evaluaties van de Commissie die de effectiviteit van wetgeving bestuderen aanzienlijk varieert. Het merendeel van deze rapporten heeft een duidelijke onderzoeksvraag en gebruikt een robuuste combinatie van consultaties met belanghebbenden en andere methoden van dataverzameling. De evaluaties doen het minder goed op andere criteria: tussen de 40% en de 70% van de rapporten presenteert een duidelijke interventielogica, heeft een valide dataverzameling en formuleert heldere conclusies. Het slechtst scoren de evaluaties op betrouwbaarheid, want slechts circa 30% van de rapporten biedt voldoende gegevens om het onderliggende onderzoek desgewenst te kunnen herhalen.

De belangrijkste verklaring voor de variatie in kwaliteit ligt bij het *type actor dat de evaluatie uitvoert*: rapporten geschreven door externe partijen in opdracht van de Commissie zijn aanzienlijk beter dan intern geproduceerde evaluaties. De gespecialiseerde expertise van externe consultants lijkt de kwaliteit van EPL evaluaties dus te verhogen. Verder laten de resultaten zien dat evaluaties van wetgeving die tot stand is gekomen met goedkeuring van het

Europees Parlement (EP) van hogere kwaliteit zijn dan andere evaluaties, al is nader onderzoek nodig om uit te zoeken waarom dit verband bestaat.

De derde voorwaarde, **systematisch gebruik**, betekent dat de resultaten van EPL evaluaties door beleidsmakers moeten worden meegewogen in hun beslissingen. Als niet aan deze voorwaarde wordt voldaan zijn de evaluaties in feite een verspilling van geld en moeite: ze kunnen alleen bijdragen aan leren en verantwoording als hun resultaten daadwerkelijk in besluitvorming worden meegenomen.

Hoofdstuk 6 van deze dissertatie laat zien dat de resultaten van de EPL evaluaties van de Commissie regelmatig gebruikt worden in impact assessments (evaluaties van de kosten en baten van Europese wetgevingsvoorstellen). Circa 65% van de impact assessments waarbij een EPL beschikbaar was verwijzen naar deze evaluatie, al variëren deze verwijzingen aanzienlijk in hun diepgang. De *tijdigheid* van de EPL evaluaties blijkt een noodzakelijke voorwaarde te zijn voor hun gebruik in impact assessments.

Hoofdstuk 7 van deze dissertatie onderzoekt het effect van politieke factoren op het gebruik van de EPL evaluaties van de Commissie voor het doel van leren. De resultaten falsificeren de hypothese dat dit gebruik afhangt van de preferenties van belangrijke actoren waar de Commissie van afhankelijk is, zoals het EP, de Raad van Ministers en grote belangengroepen. In plaats daarvan blijken vooral de *politieke prioriteiten van de Commissie zelf* grote invloed te hebben. Sinds het aantreden van de Juncker Commissie in 2014 is de institutie terughoudender geworden met het doen van nieuwe wetsvoorstellen, onder andere om het imago van de EU te beschermen tegen Eurosceptici. Vooral op beleidsterreinen die geen prioriteit zijn van de top van de Commissie is het door deze ontwikkeling lastiger geworden om de resultaten van EPL evaluaties om te zetten naar nieuw beleid. Op beleidsterreinen die wel binnen de prioriteiten van de huidige Commissie vallen is er juist veel ruimte voor EPL evaluaties om bij te dragen aan beleidsleren.

Hoofdstuk 8 van deze dissertatie behandelt het gebruik van de EPL evaluaties van de Commissie in vragen van het EP. In theorie zijn evaluatierapporten een nuttige bron van informatie voor volksvertegenwoordigers om de Commissie ter verantwoording te roepen voor haar keuzes. In de praktijk blijkt echter slechts 22% van de bestudeerde EPL evaluaties in vragen van het EP te worden aangehaald. De enige significante verklaring voor variatie op dit gebied

ligt in de *mate van conflict tussen het EP en de Commissie*: de kans dat volksvertegenwoordigers een evaluatie in hun vragen aanhalen is veel groter als het onderwerp van deze evaluatie controversieel was tijdens het wetgevingsproces.

Een belangrijke kanttekening bij alle bovenstaande resultaten is dat de Commissie als het gaat om EPL evaluaties voorop loopt in vergelijking tot veel nationale overheden. De meeste landen die lid zijn van de OESO (een organisatie die evaluatiegebruik actief stimuleert) hebben in het geheel geen systematische regels voor de initiëring, de kwaliteit en het gebruik van EPL evaluaties en de paar landen die wel systematisch zulke evaluaties uitvoeren kennen problemen die vergelijkbaar zijn aan die van de Commissie.

Algemene conclusies

Diverse academische literatuur stelt dat de Europese Commissie (deels) gedreven wordt door het belang om haar competenties te maximaliseren. Wanneer toegepast op EPL evaluaties leidt deze theorie tot de hypothese dat de initiëring en de kwaliteit van zulke evaluaties lager zijn als de Commissie een groter risico loopt dat negatieve bevindingen va zulke evaluaties kunnen leiden tot kritiek op haar competenties. De resultaten van de dissertatie bevestigen deze verwachting echter niet. Wel laten de bevindingen zien dat diverse andere politieke en technische variabelen de initiëring, de kwaliteit en het gebruik van de EPL evaluaties van de Commissie beïnvloeden. Deze factoren variëren sterk per deelonderwerp en zijn daarom hierboven reeds opgesomd.

Naast deze theoretische conclusies hebben de resultaten van deze dissertatie ook enkele praktische implicaties. Ten eerste laten de bevindingen zien dat evaluatieclausules een nuttig instrument kunnen zijn om de systematische initiëring van EPL evaluaties in de EU te bevorderen (hoewel ze geeft effect op de kwaliteit van evaluaties lijken te hebben). Ten tweede tonen de resultaten aan dat extra investeringen in evaluatiecapaciteit de Commissie kunnen helpen om een groter deel van de wetgeving van de EU te evalueren. Ten derde laten de bevindingen zien dat de tijdige beschikbaarheid van EPL evaluaties cruciaal is om hun resultaten te kunnen gebruiken in besluitvorming, wat pleit voor een strikte handhaving van het ‘evaluate first’ principe dat de Commissie heeft geformuleerd.

Al met al toont deze dissertatie aan dat de EPL evaluaties van de Commissie weliswaar een grote bijdrage leveren aan leren en verantwoording, maar dat er ook nog forse verbeteringen mogelijk zijn op het gebied van de initiëring, de kwaliteit en het gebruik van deze evaluaties. Gezien de grote invloed van Europese wetgeving op het bestaan van burgers en bedrijven valt het te hopen dat de rol van EPL evaluaties in het verbeteren van deze wetgeving in de toekomst systematischer kan worden.

Chapter 1: Introduction

Stijn van Voorst

In 2007 the European Commission, the main executive institution of the European Union (EU), initiated an evaluation of twelve EU directives on seeds and plant propagating material (the S&PM legislation). Since the 1960s these directives had sought to increase the safety of seeds by regulating their testing and marketing. However, their effectiveness had never been studied: it was unclear to what extent the legislation actually contributed to seed safety. This changed when the Commission received signals from seed producers that the implementation of the directives was causing problems: in some member states seed quality was tested extensively, whereas in others this was not the case. These signals led to the evaluation in 2007, which aimed to assess how the directives could be improved.

To enhance its quality, the evaluation was outsourced to a group of consultants led by Arcadia International. After conducting extensive interviews and surveys among stakeholders, the consultants delivered their report to the Commission in October 2008 (Arcadia International et al., 2008). In many respects the evaluation was of high quality: it presented data about all member states and clearly described its research questions, conclusions and methodology. However, the evaluation had one main flaw: the response rates of its surveys were relatively low. As a result, the Commission and various non-governmental organizations (NGOs) felt that small seed producers were underrepresented in the results.

Despite this limitation, the Commission's plant health unit still decided to use the report: it translated the evaluation's recommendations into a legislative proposal, which was published in 2013. Since the aforementioned NGOs (e.g. IFOAM EU Group, 2013: 6-11) felt that this proposal did not represent their views, they lobbied against it at the European Parliament (EP). At the EP the evaluation report remained mostly unread - which, as subsequent chapters of this dissertation will show, is rather common. A combination of the lobby of the NGOs and the upcoming EP elections caused the proposal, which so closely followed the evaluation, to be rejected by an overwhelming majority of more than seven hundred votes in March 2014.

Although the Commission's plant health unit would have liked to relaunch the proposal after its rejection, by that time a new College of Commissioners had entered into office and the topic was no longer a priority. As a result, after a process of more than six years, the S&PM legislation remained entirely unchanged (see chapter 7 for more details about this case).

The S&PM evaluation is just one instance of an ex-post legislative (EPL) evaluation conducted or outsourced by the European Commission. Essentially, such EPL evaluations are empirical studies that assess the functioning of existing EU legislation (Fitzpatrick, 2012: 479; European Commission, 2015: 271). In theory, they are supposed to contribute to the EU's 'better regulation agenda', by encouraging the improvement of legislation on the basis of objective knowledge (European Commission, 2015: 263; Fitzpatrick, 2012: 479; Luchetta, 2012: 564). By producing data about if and why legislation achieves its objectives, EPL evaluations can be a useful source of information for policy makers to decide if and how this legislation is to be amended or repealed (Fitzpatrick, 2012: 479; Vedung, 1997: 109).

The S&PM evaluation exemplifies the potential problems with the initiation, quality and use of the Commission's EPL evaluations that may limit their contributions to such legislative improvement. Concerning initiation, the Commission's reasons to launch an evaluation at a specific moment in time are sometimes illogical or unclear. Regarding quality, the consultants that usually conduct the evaluations may not deliver research that is methodologically sound. Concerning use, even if the responsible units within the Commission decide to use an evaluation, their proposals may be blocked by other institutions in the legislative process, which often do not have the time or do not see the need to read evaluation reports. The results of EPL evaluations may also be contested by interest groups and other actors in society that are affected by the legislation.

In part due to such problems with the initiation, quality and use of EPL evaluations, there is a broad variety of evaluation practices in the Commission. Some EU legislation is evaluated by the institution after just a few years, while other legislation is evaluated only after decades or not at all. There are evaluations of EU legislation that merely summarise some stakeholder opinions, whereas others base their conclusions on a broad range of data. The Commission uses some EPL evaluations to fill in every detail of new legislative proposals, while it shelves other reports immediately after their publication. This dissertation presents the first

large-scale academic effort to describe and explain such variation in the initiation, quality and use of the Commission's EPL evaluations, with the aim of assessing to what extent the Commission's system for these evaluations is fit to contribute to learning and accountability.

Section 1 of this introduction outlines the three main research questions of this dissertation, which are closely related to the three issues described above. The scope and key concepts of the research are discussed in section 2, whereas section 3 addresses the academic and practical relevance of the Commission's EPL evaluations. Section 4 and 5 of this introduction proceed with a preview of the main theoretical arguments and methodologies used throughout this dissertation. Section 6 concludes with a description of the contributions of various co-authors to the research that was conducted for this dissertation.

1. Research questions

As was explained above, EPL evaluations theoretically have an important role to play in informing decision-making about legislation. By producing knowledge about how legislation functions in reality, evaluations can be used to decide if and how such legislation should be amended or repealed (Fitzpatrick, 2012: 479; Vedung, 1997: 109). EPL evaluations may also generate knowledge about how legislation is implemented (Coglianese, 2012: 11; Vedung, 1997: 102-8). This in turn allows the actors that implement legislation - which are the Commission and national authorities in the case of the EU - to be held accountable for their actions (Højlund, 2014: 444; 2015: 35; Smith, 2015: 100; Summa and Toulemonde, 2002: 409; European Commission, 2007: 3; 2013: 2; 2015: 7).

For EPL evaluations to properly fulfil these functions of learning and accountability, an organization like the Commission must meet (at least) three requirements, each of which underpins one of the key research questions of this dissertation. The first requirement is that EPL evaluations must be *systematically initiated.* This means that in principle, all major legislation should be evaluated periodically and any exceptions or delays in this regard should be explained transparently. If EPL evaluations are not systematically initiated they can only lead to the improvement of specific laws and cannot enhance legislative quality as a whole (OECD, 2015: 120). Furthermore, if EPL evaluations are conducted selectively, the image could arise

that the Commission decides to conduct evaluations based on political considerations (Radaelli and Meuwese, 2010: 146). This, in turn, could harm the credibility of further evaluations.

Since 2007 the official procedures of the Commission prescribe that all major EU legislation should be evaluated periodically (European Commission, 2007: 22; 2015: 257). In reality, however, the Commission does not seem to live up to this promise. According to the Commission's own numbers, in 2013 29% of all important EU regulations had been evaluated, with a further 13% of such regulations being evaluated at that moment, 19% of such regulations having a future evaluation planned and no numbers being provided for directives (European Commission, 2013: 13). These numbers suggest that the Commission does not fully meet the requirement of systematic initiation: apparently, it prioritizes some pieces of legislation over others when deciding to launch EPL evaluations.

However, since these numbers only concern regulations, date back to 2013 and are not backed up by publicly available data, there is a need for a more complete, up-to-date and transparent investigation of the initiation of EPL evaluations by the Commission. Another open question is why the Commission prioritizes some EPL evaluations over others. This dissertation aims to fill these gaps in our knowledge by studying if and why there is variation in the initiation of the Commission's EPL evaluations. More formally, the first research question of this dissertation reads:

Research question 1: How can the variance in the initiation of ex-post legislative evaluations by the European Commission be explained?

Chapter 2 of this dissertation briefly answer this question in a descriptive way. Chapter 4 answers the question more extensively in both a descriptive and an explanatory way.

A second requirement for an organization to benefit from EPL evaluations is *high quality* (OECD, 2015: 121). Since evaluations are a form of applied research (Pawson and Tilley, 1997: p. xiii), they are supposed to meet criteria of methodological quality (OECD, 2015: 121). If evaluations do not meet these criteria the knowledge that they produce may be false or incomplete, which potentially leads to wasted resources and undermines the legitimacy of evaluations as a tool for policy improvement (Mayne and Schwartz, 2005: 1; OECD, 2015: 121).

To enhance the quality of both its ex-ante and its ex-post evaluations, the Commission has produced extensive guidelines that its civil servants must follow when supervising or conducting evaluations (European Commission, 2007: 2015). However, academic research has shown that the quality of the Commission's ex-ante legislative evaluations (impact assessments) varies greatly (Lee and Kirkpatrick, 2004: 17-20; Renda, 2006: 62-6; Cecot et al., 2008: 412-6), a finding that has been confirmed by the Commission's internal Regulatory Scrutiny Board (2017: 12-5). Frequent issues with the quality of impact assessments are vague problem definitions and an overreliance on subjective data (Regulatory Scrutiny Board, 2017: 13). As was shown by the case of the S&PM evaluation, methodological problems may also limit the quality and credibility of the Commission's EPL evaluations. However, so far no hard conclusions could be drawn about this subject because there has been no research about the quality of the Commission's EPL evaluations. This dissertation seeks to fill this gap in our knowledge by studying to what extent the Commission's EPL evaluations differ in quality and how these differences can be explained. In other words, it answers the following question:

Research question 2: How can the variance in the quality of ex-post legislative evaluations by the European Commission be explained?

Chapter 2 of this dissertation briefly answer this research question in a descriptive way. Chapter 5 answers the question more extensively in a descriptive and an explanatory way.

A third requirement for an organization to benefit from EPL evaluations is *systematic use* (Mayne, 2014). Even if the Commission manages to consistently produce high-quality EPL evaluations, their results still need to be considered by decision-makers to be able to contribute to aims like learning and accountability (Højlund, 2014).

Existing research shows that there is much variation in the extent to which actors use evaluations, both in general (e.g. Cousins and Leithwood, 1986; Johnson et al., 2009) and in the context of the EU (e.g. De Laat and Williams, 2014; Højlund, 2014; Højlund, 2015; Borrás and Højlund, 2015). Whereas some evaluation findings are used extensively by decision-makers to adapt policies or to improve their implementation, others remain unused. The S&PM evaluation described above is an example of an EPL evaluation that recommended significant changes to

EU directives, yet did not result in such amendments in the end. This dissertation provides a first overview of and explanation for the extent to which the Commission's EPL evaluations are used in practice. More formally, the third research question of this dissertation reads:

Research question 3: How can the variance in the use of the Commission's ex-post legislative evaluations be explained?

Chapter 6, 7 and 8 of this dissertation answer this research question in various ways. Chapter 6 addresses the use of EPL evaluations by the Commission quantitatively, while chapter 7 studies this topic in a qualitative way. Chapter 8 addresses the use of the Commission's EPL evaluations by the EP.

The third chapter of this dissertation is the only one that provides no direct answer to any of the research questions described above. Instead, this chapter measures and explains the variation in the Commission's evaluation capacity, which is an important theoretical explanation for variation in the initiation, quality and use of EPL evaluations (Nielsen et al., 2011: 325; Pattyn, 2014: 348). Therefore, chapter 3 indirectly contributes to answering all three of the main research questions of this dissertation.

The main aim of this dissertation is to answer the three research questions described above for the sake of contributing to academic knowledge. Besides this, the results of this dissertation should also result in recommendations for how the EU institutions can improve the practice of EPL evaluations. These recommendations are provided in the final conclusion of this dissertation.

2. Definitions and scope

This dissertation defines an ex-post legislative evaluation (EPL evaluation) as an empirical study initiated by the European Commission that retrospectively assesses the functioning of generally binding EU legislation. In this definition, 'functioning' can refer to a broad number of criteria, such as the implementation, cost-benefit ratio or (un)intended effects of the evaluation's object of study.

The concept of 'generally binding European legislation' refers to EU regulations, directives and treaty articles. Evaluations of decisions about single cases or non-binding rules therefore fall outside of the scope of this dissertation. Although such evaluations could be an interesting topic for academic research, they differ from EPL evaluations as defined above because the Commission's better regulation agenda and evaluation guidelines do not fully apply to them (European Commission, 2015: 35, 73). Moreover, evaluations of single decisions are unlikely to result in general policy changes and evaluations of non-binding rules cannot result in enforcement actions. Therefore, such evaluations can be expected to be driven by different mechanisms than EPL evaluations.

It should be noted that the definition provided above deviates from the official description of EPL evaluations used by the Commission. Since 2015, the Commission considers an evaluation to be a staff working document that assesses the effectiveness, efficiency, relevance, coherence and EU added value of a policy (European Commission, 2015: 271, 289). Although such documents are often based on reports by external consultants that may also bear the title 'evaluation', these are not officially recognized as such by the Commission. Reports that only assess some of the criteria listed above or only assess the implementation of legislation are also not considered full evaluations by the Commission. Instead, they are referred to as 'studies' or simply 'reports'.

There are three reasons to deviate from the Commission's official definition in this dissertation. Firstly, the data presented here mostly concern the years 2000-2014, which was before the Commission clarified its definition of an 'evaluation' in 2015. Secondly, because evaluations and 'studies' or 'reports' about legislation often differ solely in the number of aspects that they assess, there is no reason to assume that different criteria should apply to their initiation, quality or use, or that these issues should be studied with different methods. Thirdly, an exclusive focus on the Commission's official 'evaluations' would limit the number of cases that can be studied, which would negatively affect the causal validity and external validity of this dissertation. Out of the 313 cases included in the main dataset used for this dissertation (which is further described below), 99 cases bear the official title 'evaluation', 56 cases bear the official title 'study', 120 cases bear the official title 'report' and the approximately fifty other cases[1] bear a variety of other titles, like 'assessment' or 'appraisal'.

The reason why this dissertation focuses on EPL evaluations of the Commission is that it is the leading executive organization of the EU and therefore bears the main responsibility for evaluating European policies (Stern, 2009: 70-1; EC, 2015: 253). For this reason, the Commission is the only EU institution that can be expected to initiate and use EPL evaluations on a large scale. Indeed, the other main institutions of the EU only conduct evaluations to a limited degree. Whereas the EP has had a research service that may conduct EPL evaluations since 2012 (European Parliamentary Research Service, 2017: 7), in June 2017 this service had only conducted 33 ex-post evaluations in total.[2] The Council and the European Council had no permanent services for ex-post evaluations at the time this dissertation was completed. The European Court of Auditors has produced some performance audits, meta-evaluations and special reports that assess EU legislation indirectly, but usually it does not evaluate individual pieces of legislation (Stephenson, 2015). The few EPL evaluations that have been conducted by these institutions could also be driven by different factors than the Commission's evaluations, which makes it appropriate to exclude them from this dissertation.

Now that the definition of an EPL evaluation as used in this dissertation has been clarified, the question arises why this topic is worthy of academic scrutiny. The next section therefore discusses the relevance of the Commission's EPL evaluations from both a theoretical and a practical perspective.

3. Relevance

Because of its strong reliance on legislative policies, the EU has often been dubbed a 'regulatory state' (Lodge, 2008: 282; Majone, 1999: 1; Radaelli, 1999: 759). As the Commission plays a central role in these policies, most of its legislative tasks have received ample academic scrutiny. For example, many scholars have studied the Commission's role in initiating legislative proposals, producing delegated and implementing legislation and enforcing national compliance with EU legislation (for an overview of relevant literature, see Kassim et al., 2013; McCormick, 2015: 155-74; Schmidt and Wonka, 2013; Wille, 2013).

In contrast to this extensive attention for the Commission's tasks to produce and enforce legislation, the institution's role in the final stage of the EU legislative process has largely been ignored: there is very little academic work about EPL evaluations in the EU. While

various authors have paid attention to ex-post evaluations of EU spending programmes (e.g. Bachtler and Wren, 2006; Baslé, 2007; Højlund, 2014; Borrás and Højlund, 2015) and impact assessments of proposals for new EU legislation (e.g. Cecot et al., 2008; Radaelli, 2009; Radaelli and Meuwese, 2010; Torriti, 2010), EPL evaluations in the EU have received very little academic scrutiny. The exception to this are general texts about evaluation in the EU that include some paragraphs about EPL evaluations (e.g. Højlund, 2015; Summa and Toulemonde, 2002; Stame, 2008; Stern, 2009), articles that discuss the Commission's EPL evaluations as a form of input for impact assessments (e.g. Luchetta, 2012; Smismans, 2015), and a paper about such evaluations written by a practitioner (Fitzpatrick, 2012).

This lack of attention is all the more surprising given the theoretical importance of EPL evaluations. As explained above, EPL evaluations may produce knowledge about the effectiveness and implementation of legislation, thus making them a potential source of information for the Commission and other decision-makers when proposing policy changes (Fitzpatrick, 2012: 479; Vedung, 1997: 102-9). By doing so, EPL evaluations are both the final step in the EU's legislative process and a potential first step in a process of amendments (Smismans, 2015: 19).

Besides this theoretical relevance, EPL evaluations are also increasingly important for the day-to-day activities of the Commission. The institution first emphasized the importance of EPL evaluations for legislative improvement and accountability in 2000, after which it started to systematize its procedures for such evaluations from 2007 onwards (Fitzpatrick, 2012: 478; European Commission, 2007: 3-4). Since 2010 the Commission has also stressed the role of EPL evaluations in judging the suitability of entire regulatory frameworks (so-called 'fitness checks') (European Commission, 2010: 5). Furthermore, from 2012 onwards it has given EPL evaluations a central place in its REFIT programme, which aims to identify and remove superfluous rules (European Commission, 2012: 4). In 2015 the Commission published a new 'better regulation toolbox' that included extensive guidelines for EPL evaluations (European Commission, 2015). This was a significant development because the Commission's previous evaluation guidelines mostly focused on spending programmes (European Commission, 2004).

Given the theoretical importance of EPL evaluations and the increased attention that they receive in practice at the EU level, it is important for scholars to critically assess if and why

the Commission engages in evaluated-related activities. Does the Commission indeed systematically initiate and use high-quality evaluations? Is the purpose of the Commission's evaluation-related activities really to improve learning and accountability, or do other motives inform these efforts? To answer these questions and more, this dissertation presents a first academic effort to systematically describe and explain the initiation, quality and use of the Commission's EPL evaluations.

4. Theoretical framework

Political and technical explanations

Despite the existence of a vast literature about evaluation methods and techniques (e.g. Vedung, 1997; Nielsen et al., 2011; Rossi et al., 2004), there is a lack of comprehensive explanatory theories about the initiation, quality and use of evaluations. However, empirical research has revealed various individual factors that may explain these phenomena. These factors can broadly be divided in two categories: political and technical explanations (Bovens et al., 2008: 120; Schwartz, 1998: 295; Weiss, 1993: 94).

Political explanations, firstly, refer to the interests that actors have in (not) conducting evaluation-related activities like initiating an evaluation, investing in evaluation quality and using evaluation results. The logic behind these explanations is that evaluation-related activities are inherently subjective: they will be supported by actors to which they are advantageous and opposed by actors to which they are disadvantageous (Bovens et al., 2008: 120; Schwartz, 1998: 295; Vedung, 1997: 111; Weiss, 1993: 95-8).

Following this logic, political actors like the EU institutions can be expected to engage in evaluation-related activities that are beneficial to them and to refrain from evaluation-related activities for which the opposite is the case. This expectation is especially plausible for EPL evaluations, as legislative changes are always discussed in parliament and are therefore likely to receive much attention at the political level (Bussmann, 2014: 1; Højlund, 2015: 45). The S&PM evaluation that was discussed at the beginning of this introduction is an example of an evaluation that failed to change existing policies due to political considerations in the European Parliament, such as upcoming elections and active opposition by NGOs.

'Technical' explanations, secondly, are related to the capacity and formal obligations to conduct evaluations. The logic behind these explanations is that some evaluations have to be prioritized over others due to limited resources. Therefore, it can be expected that organizations that invest more human and financial capital in evaluations will initiate more and better EPL evaluations and make more use of their results (Nielsen et al., 2011: 325; Pattyn, 2014: 348). It can also be expected that organizations will prioritize investing in evaluations that are made compulsory by either general procedures or evaluation clauses in specific pieces of legislation (Summa and Toulemonde, 2002: 410).

Concerning evaluation use, slightly different expectations are formulated throughout this dissertation. Because decisions concerning legislative changes are always at the discretion of the legislator, the use of EPL evaluations is never made compulsory by evaluation clauses. In EU legislation such clauses may prescribe when and sometimes how the Commission must conduct an EPL evaluation, but not whether it should implement the results. Therefore, in this dissertation the presence of evaluation clauses is not expected to affect evaluation use. Conversely, a factor that is expected to influence evaluation use in particular is evaluation quality: decision-makers are more likely to use evaluations when they trust that their results are robust (Johnson et al., 2009: 377-378; De Laat and Williams, 2014: 158-67).

These political and technical explanations for evaluation-related activities can be linked to different views about the nature of the European Commission. The Commission used to be viewed as a technocratic institution that fulfilled the tasks that the EU's member states delegated to it to the best of its abilities (Radaelli, 1999: 759; Wille, 2013; Franchino, 2007: 11; Boswell, 2008: 472; Hartlapp et al., 2014: 1). In this context, 'technocratic' refers to decision-making on the basis of objective knowledge (Radaelli, 1999: 759), for which EPL evaluations may be a useful tool (Fitzpatrick, 2012: 479). Over time the technocratic perspective has been replaced by the view that the Commission is (also) a political institution that pursues its own preferences (e.g. Cini, 2015; Franchino, 2007: 11; Wille, 2013: 191-192; Wonka, 2015), for example by protecting and increasing its competences (Hartlapp, 2014: 1-14; Majone, 1996: 65; Nugent and Rhinard, 2016: 1201; Tallberg, 2003: 28). If the Commission indeed operates as a political institution we would expect political factors to best explain its evaluation-related decisions, whereas if it operates as a technocratic institution we would expect technical factors

to be more important. Therefore, this dissertation also aims to shed some further light on the nature of the Commission and its role in EU governance.

Application per chapter

The political and technical explanations described above are applied in various ways throughout this dissertation, depending on the specific content of each chapter. Chapter 2 is descriptive in nature and therefore does not have an explanatory theoretical framework. However, this chapter's conclusion does highlight the potential of political and technical explanations for the initiation and quality of evaluations.

Chapter 3 provides three possible explanations for variation in the capacity of the Commission's directorates-general (DGs) to conduct EPL evaluations: the amount of legislation for which a DG is responsible, the presence of a tradition of evaluating spending programmes and the sensitivity of a DG's policy field. The first two of these explanations are technical in nature, because they focus on the extent to which DGs must build evaluation capacity due to their legislative obligations and the extent to which they have the experience needed to do so. The third explanation is political in nature, as it predicts that DGs with policy fields that are politically sensitive build less evaluation capacity, since for them evaluation results may be particularly threatening.

Chapter 4 presents two motives for the Commission to (not) initiate an EPL evaluation: an enforcement motive and a strategic motive. Both of these motives are political in nature, as they concern the potential advantages and disadvantages of EPL evaluations to the Commission's interests. On the one hand, EPL evaluations may be useful for the Commission to check legislative implementation by the member states (enforcement motive), while on the other hand they may threaten the Commission's competences if their findings are negative (strategic motive). Technical explanations like the presence of evaluation clauses and the evaluation capacity of the responsible DGs are treated as control variables in this chapter.

Chapter 5 focuses on the quality of the Commission's EPL evaluations in relation to their suitability to learn about legislative effectiveness. Therefore, factors related to the role of such evaluations in enforcement processes were omitted here. However, the strategic motive presented in chapter 4 is also considered in this chapter, as the Commission may have an

incentive to distort the quality of EPL evaluations when there is a risk that negative findings could threaten its competences. The effects of technical variables like the evaluation capacity of the responsible DGs, the type of evaluator and the complexity of the evaluated legislation are also assessed in this chapter.

Chapter 6 studies three technical explanations for the use of the Commission's EPL evaluations in subsequent impact assessments (and vice versa): the timeliness of the EPL evaluations, their overall quality and their scope. All of these explanations are related to the practical possibilities to use an evaluation in an impact assessments (and vice versa), which is expected to be difficult if an EPL evaluation is not available on time or if it does not provide the required information. Political explanations for use were not considered in this chapter because of a lack of reliable quantitative indicators for such variables.

Conversely, chapter 7 focuses specifically on the influence of political factors on the use of EPL evaluations by the Commission. In this chapter, technical explanations for use were held constant by studying three cases that were all of high quality and were all conducted by the same DG. The central theoretical expectation of this chapter is that the absence of opposition to an evaluation's findings by important political actors is a necessary condition for use. In other words, if the Commission, the EP, the Council or all major interest groups oppose a recommendation provided by an EPL evaluation, we can expect this recommendation to remain unused when subsequent legislative proposals are drafted.

Chapter 8 provides four possible explanations for variation in the use of the Commission's EPL evaluations by members of the EP (MEPs): the objectivity of the evaluation's results, the communicative quality of the evaluation, the salience of the evaluation's topic to MEPs and the level of conflict during the legislative process. The first two of these explanations are technical[3] in nature because they are dimensions of evaluation quality - the chances that an EPL evaluation is used by MEPs can be expected to increase if they trust its results. The other two explanations are political in nature, as they concern MEPs' interests in controlling the way in which the Commission implements policies.

5. Methods and data

Data collection and case selection

Before the research presented in this dissertation was conducted, no large-scale overview of the Commission's EPL evaluations existed. Therefore, a unique dataset of 313 EPL evaluations was constructed for the purpose of this dissertation. The evaluations were collected from a large number of sources, including various webpages and reports of the Commission as well as the EU Bookshop, Eur-lex, and systematic Google searches. For a full overview of the dataset and its sources, see chapter 2 and four of this dissertation.

The use of this dataset enhances the external validity of the dissertation: most of the research findings represent (almost) the entire population of publicly available EPL evaluations, at least within the timeframe of the data collection. This timeframe differs somewhat between the chapters. In chapter 2 and eight the dataset includes about 220 evaluations from 2000-2012, as these chapters were completed during 2014. The other chapters were written during 2015-2017 and are therefore based on the 'full' dataset of 313 EPL evaluations from 2000-2014. The timeframe per chapter is summarised in Table 1.

The reason for only including evaluations published since 2000 is that the Commission formulated the ambition to systematically evaluate EU legislation for the first time during that year (Fitzpatrick, 2012: 478). Furthermore, evaluations published before 2000 are less likely to have been published online. The reason to end the data collection at 2014 is that it often takes some time before all evaluations from a certain year are published. Therefore, if EPL evaluations from 2015-2017 had been studied as well, there would likely have been gaps in the data collection for these years. Such gaps could have led to biases in the results.

Three additional datasets were used for specific chapters of this dissertation. Firstly, chapter 3 uses a dataset of seventeen directorates-general (DGs) of the Commission to explain variance in their evaluation capacity. Secondly, chapter 4 uses a dataset of 277 major pieces of EU legislation from 2000-2004 to study why some of this legislation is evaluated while other legislation is not. An initial version of this dataset is also used in chapter 2. Thirdly, chapter 6 uses a dataset of 225 impact assessments to study the extent to which EPL evaluations feed into subsequent legislative processes. Table 1 provides a full list of these additional datasets.

Table 1: overview of research methods

Chapter number	**Topic**	**Datasets used**	**Method of data collection**	**Method of analysis**
2	Description of dataset	Dataset of 216 EPL evaluations 2000-2012	Quantitative document analysis	Descriptive analysis
3	Evaluation capacity	Dataset of 17 DGs dealing with legislation	Interview & qualitative document analysis	QCA
4	Initiation of evaluations	Dataset of 313 EPL evaluation 2000-2014 & Dataset of 277 major pieces of EU legislation 2000-2004.	Quantitative document analysis	Binary logistic regression
5	Evaluation quality	Dataset of 313 EPL evaluation 2000-2014.	Quantitative document analysis	Linear regression
6	Evaluation use	Dataset of 313 EPL evaluation 2000-2014 & dataset of 225 impact assessments 2003-2014.	Quantitative document analysis	QCA
7	Evaluation use	Dataset of 313 EPL evaluation 2000-2014.	Interviews & qualitative document analysis	Process tracing
8	Evaluation use	Dataset of 220 EPL evaluations 2000-2012	Quantitative document analysis	Binary logistic regression

Although most of the research presented in this dissertation is quantitative, case studies were used as well for chapter 7. Their main purpose was to delve into the underlying mechanisms of the use of EPL evaluations: *why* do certain variables explain variance in such use? The full dataset of EPL evaluations was used to select appropriate cases for this endeavour, thus mixing a quantitative and a qualitative approach.

Methods of analysis

In chapter 4, 5 and 8 of this dissertation various forms of regression analysis are the main method of analysis, as this is the most suitable technique to answer explanatory research questions based on quantitative data (Field, 2013: 768-810; Long, 1997: 42). The other chapters use a variation of other methods of analysis. Chapter 2 is entirely descriptive and therefore features no explanatory analysis. Chapter 3 and 6 are based on quantitative datasets that are too small or have too few positive scores on their dependent variables to make regression analysis viable. Therefore, QCA was used as the method of analysis for these chapters, as this technique can be used with small numbers of cases. An additional advantage of QCA is that it allows for studying combinations of factors that may explain a certain outcome (Ragin, 2008: 9).

Chapter 7 is entirely based on in-depth case studies and therefore features process-tracing as its method of analysis: the detailed examination of sequences of events to study if the causal mechanisms implied by a certain theory are indeed present (George and Bennett, 2005: 9). Table 1 summarises the method of analysis and the other methodological characteristics of each chapter.

6. Articles and co-authorships

The seven substantive chapters of this dissertation (chapter 2-8) were originally written as individual articles. At the time this dissertation was completed, six of these articles had been published in peer-reviewed journals; the full references to these publications can be found at the beginning of the corresponding chapters. The texts of these articles are identical to the texts of the associated chapters, although their lay-out and references have been updated to make them look more consistent.[4] At the time of writing, only the article about evaluation

quality that underlies chapter 5 had not yet been accepted for publication. The content of the final version of this article may therefore deviate from the corresponding chapter if it is revised during its review process.

Out of the seven substantive chapters, chapter 3 was written without any co-authors. The other six chapters include at least some contributions from other academics. For the sake of transparency these contributions are listed below. All co-authors have given explicit approval to include the articles that they have helped to produce in this dissertation.

Concerning chapter 2, Prof. Dr. Ellen Mastenbroek is the first author and Prof. Dr. Anne Meuwese is the third author. My main contribution as second author was constructing the dataset of EPL evaluations that is presented in this chapter (and is also used in all other chapters of the dissertation) under the supervision of Prof. Dr. Ellen Mastenbroek. I also wrote most of the methodology and results sections of this chapter and I assisted in writing the other parts of the text; the co-authors wrote most of the other sections of this chapter.

Chapter 4 and 5 are joint publications with Prof. Dr. Ellen Mastenbroek as the second author. Her main contributions to these chapters were developing an initial version of the theoretical framework and providing feedback throughout the research and writing process.

Chapter 6 is a joint publication with Thomas van Golen LLM MSc. The work conducted for this study was split equally between both authors and the order of their names on the publication was therefore determined alphabetically. Thomas van Golen LLM MSc collected all of the data about impact assessments that is presented in this chapter and wrote the parts of the text that concern the use of EPL evaluations by impact assessments. Conversely, I collected all of the data about EPL evaluations that is presented in this chapter and wrote the parts of the text that concern the use of impact assessments by EPL evaluations.

Chapter 7 is a joint publication with Dr. Pieter Zwaan as the second author. His main contributions to the chapter were developing and drafting parts of the research methodology, providing extensive feedback throughout the research process and providing assistance during three interviews.

Regarding chapter 8, Dr. Pieter Zwaan is the first author and Prof. Dr. Ellen Mastenbroek is the third author. My own contributions to this chapter mainly concerned the data collection

and analysis, writing the methodology and results sections and assisting on writing other parts of the text; the co-authors wrote most of the other parts of this chapter.

The reason to include these seven chapters in this dissertation despite their various sets of co-authors is that they all concern different steps in the process of the Commission's EPL evaluations. This makes reading them in combination with each other especially valuable. Furthermore, all seven of the chapters are in some way related to the dataset of EPL evaluations that is described in detail in chapter 2. Together, the seven chapters aim to provide a comprehensive picture of the dataset and the process of EPL evaluation in the Commission, which would not have been possible if some of them had been left out.

Besides the contributions of the various co-authors, full credit is given to Prof. Dr. Ellen Mastenbroek for setting up the project about EPL evaluations that led to this dissertation and to the Netherlands Organisation for Scientific Research (in Dutch: Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) for funding the project.[5] The assistance of these actors was crucial for the production of this dissertation.

Notes

[1] The categories mentioned here are not entirely mutually exclusive. For example, there is one case that is called an 'evaluation study' and one other case that is called an 'evaluation study report'. This is why the total number of cases mentioned in the text is not exactly 313.

[2] The number of 33 ex-post evaluations was received from the ex-post evaluations unit of the European Parliamentary Research Service (EPRS) via e-mail contact with eprs-expostevaluation@europarl.europa.eu at 26 June 2017. According to the e-mail from this unit, 23 'European implementation assessments' had been published by the research service before the summer of 2017. These assessments are essentially ex-post evaluations of the implementation of European policies. There were ten further reports categorized as 'other ex-post evaluations', adding up to 33 ex-post evaluations in total. These evaluations usually concern legislation, but not always; exact numbers in this regard could not be provided. Two other types of reports from the EPRS that may partly evaluate EU legislation are 'implementation appraisals' (64 in total) and 'rolling-check lists' (13 in total). However, these publications take the form of brief notes rather than full reports and are therefore no EPL evaluations as defined in this dissertation.

[3] In chapter 8 these two explanations are called 'rationalistic' instead of 'technical'. This is due to the fact that this chapter was published as an article in an early stage of the PhD project - for the later articles the term 'technical' has been preferred because it is less ambiguous. The meaning of both words is the same in the context of this dissertation.

[4] In particular, all references were made consistent with the APA-style used by the *Journal of European Public Policy*, in which chapter 2 and 7 of this dissertation have been published. This resulted in some significant changes to chapter 6, which had previously been published in a journal that uses a completely different style of referencing. The footnotes used in the original version of that article were changed to in-text references and a list of references

was added at the end of its text. For the other chapters the changes were relatively minor, although some mistakes in the references made in the original articles have been fixed.

[5] The official title and number of the project that was funded by the Netherlands Organisation for Scientific Research were: 'Closing the regulatory cycle? Ex-post legislative evaluation in the European Union (406-14-030)'.

References

Arcadia International, Van Dijk Management Consultants, Civic Consulting and Agra CEAS (2008) *Evaluation of the Community acquis on the marketing of seed and plant propagating material (S&PM)*. Brussels: European Commission.

Bachtler J and Wren C (2006) The evaluation of EU Cohesion Policy: Research questions and policy challenges. *Regional Studies* 40(2): 143-153.

Baslé M (2007) Strengths and weaknesses of European Union policy evaluation methods: Ex-post evaluation of objective 2, 1994–99. *Regional studies* 40(2): 225-235.

Borrás S and Højlund S (2015) Evaluation and policy learning: The learners' perspective. *European Journal of Political Research* 54(1): 99-120.

Boswell C (2008) The political functions of expert knowledge: Knowledge and legitimation in the European Union. *Journal of European Public Policy* 15(4): 471-488.

Bovens M, 't Hart P and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford handbook of public policy*. Oxford: University Press, pp. 320-335.

Bussmann W (2014) *What happens after a law gets evaluated? The interplay between program managers, the executive and the parliament*. In: ECPR Fifth Biannual Conference on Regulatory Governance, Barcelona, Spain, 25-27 June 2014.

Cecot C, Hahn RW, Renda A and Schrefler L (2008) An evaluation of the quality of impact assessment in the European Union with lessons for the US and the EU. *Regulation and Governance* 2(4): 405-424.

Cini M (2015) The European Commission - Politics and Administration. In: Bauer M and Trondal J (eds) *The Palgrave Handbook of the European Administrative System*. Houndmills: Palgrave Macmillan, pp. 127-144.

Coglianese C (2012) *Evaluating the performance of regulation and regulatory policy*. Report to the Organization of Economic Cooperation and Development.

Cousins JB and Leithwood KA (1986) Current empirical research on evaluation utilization. *Review of Educational Research* 56(3): 331-364.

De Laat B and William K (2014) Evaluation use within the European Commission: lessons for the Commissioner. In: Loud ML and Mayne J (eds) *Enhancing Evaluation Use:*

Insights from Internal Evaluation Units. London: Sage, pp. 147-174.

European Commission (2004) *Evaluating EU activities: A practical guide for the Commission services.* Brussels: European Commission.

European Commission (2007) *Communication to the Commission from Ms Grybauskaité in agreement with the President: Responding to strategic needs: Reinforcing the use of evaluation [SEC(2007)213]*. Brussels: European Commission.

European Commission (2010) *Multi-annual overview (2002-2009) of evaluations and impact assessments*. Available at: http://ec.europa.eu/smart-regulation/evaluation/docs/multiannual_overview_en.pdf (Accessed 10 July 2015).

European Commission (2012) *EU regulatory fitness [COM(2012)746]*. Brussels: European Commission.

European Commission (2013) *Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions. Strengthening the foundations of smart regulation: improving evaluation [COM(2013)686]*. Brussels: European Commission.

European Commission (2015) *Better Regulation Toolbox [SWD(2015)111]*. Brussels: European Commission.

Field A (2013) *Discovering statistics: using SPSS (and sex and drugs & rock 'n roll) (4th edition)*. London: Sage.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

Franchino F (2007) *The powers of the Union: Delegation in the EU.* Cambridge: University Press.

George AL and Bennett A (2005) *Case studies and theory development in the social sciences.* Cambridge, MA: MIT.

Hartlapp M, Metz J and Rauh C (2014) *Which policy for Europe? Power and conflict inside the European Commission*. Oxford: University Press.

Højlund S (2014) Evaluation use in evaluation systems - the case of the European Commission. *Evaluation* 20(4): 428-446.

Højlund S (2015) Evaluation in the European Commission - for accountability or learning? *European Journal of Risk Regulation* 6(1): 35-46.

IFOAM EU Group (2013) *Towards more crop diversity - adapting market rules for future food security, biodiversity and food culture*. Brussels: online publication.

Johnson K, Greenseid LO, Toal SO, King JA, Lawrenz F and Volkov B (2009) Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation* 30(3): 377-410.

Kassim H, Peterson J, Bauer MW, Connolly S, Dehousse R, Hooghe L and Thompson A (2013) *The European Commission of the Twenty-First Century.* Oxford: University Press.

Lee N and Kirkpatrick C (2004) *A Pilot Study of the Quality of European Commission Extended Impact Assessments*. Impact assessment research center.

Lodge M (2008) Regulation, the regulatory state and European politics. *West European Politics* 31(1-2): 280-301.

Long JS (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Luchetta G (2012) Impact Assessment and the Policy Cycle in the EU. *European Journal of Risk Regulation* 3(4): 561-575.

Majone G (1996) *Regulating Europe*. London: Routledge.

Majone G (1999) The regulatory state and its legitimacy problems. *West European Politics* 22(1): 1-24.

Mayne J (2014) Issues in enhancing evaluation use. In: Loud ML and Mayne J (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. London: Sage, pp. 1-14.

Mayne J and Schwartz R (2005) Assuring the quality of evaluative information. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction, pp. 1-17.

McCormick J (2015) *European Union Politics (2nd edition)*. London: Palgrave.

Nielsen SB, Lemire S and Skov M (2011) Measuring evaluation capacity: Results and implications of a Danish study. *American Journal of Evaluation* 32(3): 324-344.

Nugent N and Rhinard M (2016) Is the European Commission Really in Decline? *Journal of Common Market Studies* 54(5): 1199-1215.

OECD (2015) *OECD Regulatory Policy Outlook 2015*. Paris: OECD Press.

Pattyn V (2014) Why organizations (do not) evaluate? Explaining evaluation activity through the

lens of configurational comparative methods. *Evaluation* 20(3): 348-367.

Pawson R and Tilley N (1997) *Realistic evaluation*. London: Sage.

Radaelli CM (1999) The public policy of the European Union: Whither politics of expertise? *Journal of European Public Policy* 6(5): 757-774.

Radaelli CM (2009) Rationality, power, management and symbols: Four images of regulatory impact assessment. *Scandinavian Political Studies* 33(2): 164-188.

Radaelli CM and Meuwese ACM (2010) Hard questions, hard solutions: Proceduralisation through impact assessment in the EU. *West European Politics* 33(1): 136-153.

Ragin CC (2008) *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University Press.

Regulatory Scrutiny Board of the European Commission (2017) *Regulatory Scrutiny Board - annual report 2016.* Brussels: European Commission.

Renda A (2006) *Impact assessment in the EU: The state of the art and the art of the state*. Brussels: Centre for European Policy Studies.

Rossi PH, Lipsey MW and Freeman HE (2004) *Evaluation: A systematic approach (7th edition)*. Thousand Oaks, CA: Sage.

Schmidt SK and Wonka A (2013) The European Commission. In: Jones E, Menon A and Weatherill S (eds) *The Oxford Handbook of the European Union*. Oxford: University Press, pp. 336-349.

Schwartz R (1998) The Politics of Evaluation Reconsidered: A Comparative Study of Israeli Programs. *Evaluation* 4(3): 294-309.

Smismans S (2015) Policy Evaluation in the EU: The Challenges of Linking Ex Ante and Ex Post Appraisal. *European Journal of Risk Regulation* 6(1): 6-26.

Smith M (2015) Evaluation and the Salience of Infringement Data. *European Journal of Risk Regulation* 6(1): 90-100.

Stame N (2008) The European project, federalism and evaluation. *Evaluation* 14(2): 117-140.

Stephenson P (2015) Reconciling Audit and Evaluation? The Shift to Performance and Effectiveness at the European Court of Auditors. *European Journal of Risk Regulation* 6(1): 79-89.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK,

Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 67-85.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation*. New Brunswick: Transaction Publishers, pp. 407-424.

Tallberg J (2003) *European governance and supranational institutions: Making states comply*. Abingdon, Oxon: Routledge.

Torriti J (2010) Impact assessment and the liberalization of the EU energy markets: Evidence-based policy-making or policy-based evidence-making? *Journal of Common Market Studies* 48(4): 1065-1081.

Vedung E (1997) *Public policy and program evaluation.* New Brunswick: Transaction.

Weiss CH (1993) Where Politics and Evaluation Research Meet. *American Journal of Evaluation* 14(1): 93-106.

Wille A (2013) *The normalization of the European Commission: Politics and bureaucracy in the EU executive*. Oxford: University Press.

Wonka A (2015) The European Commission. In: Richardson J and Mazey S (eds) *European Union. Power and policy-making (4th edition)*. London: Routledge, pp. 83-105.

Chapter 2: Closing the regulatory cycle? A meta evaluation of ex-post legislative evaluations by the European Commission

Ellen Mastenbroek, Stijn van Voorst and Anne Meuwese

Published as: Mastenbroek E, Van Voorst S and Meuwese A (2016) Closing the regulatory cycle? A meta-evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy* 23(9): 1329-1348.

Abstract

Theoretically, ex-post legislative (EPL) evaluations play an important role in the European regulatory cycle. By critically assessing the administration, compliance or outcomes of legislation, they may allow for learning and inform enforcement. At the same time, the European Commission may have incentives *not* to evaluate, as EPL evaluations may lead to undesired policy change or repeal. Furthermore, the development of systematic, high-quality EPL evaluations is threatened by more technical problems in the sphere of evaluability. Hence, the odds are against the systematic production of high-quality evaluations in the European Union (EU). This article assesses this argument by conducting a meta evaluation of the coverage and quality of ex-post legislative evaluations by the European Commission, using two novel datasets. The main findings are that EPL evaluation coverage indeed is patchy, with no clear upward trend in recent years. EPL evaluation is primarily a matter of legislative obligation instead of own initiative. There is great scope, finally, for enhancing the quality of EPL evaluations, by improving methodological quality, stakeholder involvement and transparency.

1. Introduction

'Evaluation answers the question of whether a treatment works in terms of reducing a problem' (Coglianese, 2012: 14). By critically assessing the administration, compliance or outcomes of legislation, ex-post legislative (EPL) evaluation should enable policy-makers to establish faults in

the linkages between law-on-paper and actual effects (Coglianese, 2012: 14). By providing a learning opportunity (Smismans, 2015: 12) and a basis for enforcement (Stame, 2008: 124), EPL evaluations may play an important role in the regulatory cycle.

The European Commission (EC) (2000, 2001, 2004, 2010a, 2013, 2015[1]), clearly recognizing the potential of EPL evaluations, has repeatedly committed itself to systematic, high-quality EPL evaluation. It has offered a range of rationales for doing so, including priority-setting and resource allocation (European Commission, 2004: 9), enhancing effectiveness (European Commission, 2001: 10), accountability (European Commission, 2010a: 2), legitimacy (European Commission, 2007: 3) and enforcement (European Commission, 2010a: 7). EPL evaluations have come to constitute a key building block of the Commission's Smart Regulation strategy, and its current Regulatory Fitness and Performance (REFIT) programme, according to which evaluations should inform legislative review (Smismans, 2015: 11-12; European Commission, 2013: 3; 2014: 13-14).

Despite their theoretically important role in the regulatory cycle, the systematic production and high quality of EPL evaluations are not guaranteed. In line with both the political and rationalistic view of evaluation (Bovens et al., 2008), two obstacles can be identified. First, because EPL evaluations may uncover critical problems in the actual working of legislation, they may lead to calls for legislative repeal. This potential outcome clashes with the Commission's alleged strategy of continuous legislative expansion (Lodge, 2008: 286; Majone, 1999: 2) and creates a risk of selective, biased or even absent evaluations. This way, the Commission, seen as an agent of the Council, may try to avoid control by the Council (Pollack, 1997: 109; Tallberg, 2003: 19). In its role as guardian of the treaties, however, the Commission may use evaluations to uncover faults in member state implementation. Second, according to the rationalistic view, systematic high-quality evaluation may be hampered by problems of evaluability (Fitzpatrick, 2012: 480; Summa and Toulemonde, 2002). In sum, both the political and rationalistic perspective on evaluation predict that, despite its theoretical potential, systematic high-quality EPL evaluation by the Commission is not likely to materialize, with the exception of evaluations aimed at establishing national implementation.

This article seeks to evaluate the coverage and quality of EPL evaluation through a meta evaluation of the EU's evaluation system (Cooksy and Caracelli, 2005), focusing on its main

outputs (Schwartz and Mayne, 2005: 6): specific EPL reports. This meta evaluation is embedded in an analytical framework comprising four aspects of evaluation coverage (proportion, number, obligatory character and object of evaluation) and four quality variables (evaluation type, methodological quality, process quality and usefulness). The meta evaluation is based on two novel datasets compiled for this purpose: one containing 216 EPL evaluations: one containing 156 major EU directives and regulations.

By evaluating the EU's evaluation system, this article paves the way for future theory-testing studies on this virtually neglected stage in the EU regulatory cycle. It thus seeks to bring EPL legislative evaluation[2] on a par with, firstly, the well-researched other stages of the EU regulatory cycle (Versluis et al., 2011), and, secondly, other types of EU evaluation, most notably impact assessment (Renda, 2006; Radaelli, 2009; Radaelli and Meuwese, 2010) and programme evaluation (Bachtler and Wren, 2006).

2. EPL evaluations in the EU regulatory cycle

EPL evaluation in the EU is the province of the European Commission (Poptcheva, 2013: 2) or, rather, individual DGs within it (Summa and Toulemonde, 2002: 413). The EU's Financial Regulation (966/2012) stipulates that results of periodic evaluations of community actions should be taken into account in budgetary decisions (Art. 30). Accordingly, ex-post evaluations of expenditure programmes have been conducted from the 1980s, to the point where this became well-established practice (Fitzpatrick, 2012: 478). EPL evaluations, however, have been much less common (Fitzpatrick, 2012: 479).[3] Therefore, in 2000 the Commission committed itself to conduct evaluations on legislation with substantial impacts (Poptcheva, 2013: 2). The 2001 Mandelkern report triggered the EU's Better Regulation policy, which seeks to improve the quality of EU legislation through regulatory impact assessment, consultation standards and simplification programmes (Radaelli and Meuwese, 2009: 640).

The Commission's 2001 White Paper on Governance called for more evidence-based decision making (Poptcheva, 2013: 2). It established effectiveness as an important principle of good governance, arguing that more legislative evaluation clauses were required. Yet, this did not increase attention for EPL evaluation. The Commission's 2002 Communication on Better Lawmaking (European Commission, 2002) confirmed systematic ex-ante impact assessment as

the tool for enhancing effectiveness. In 2007, the tide seemed to turn as the Commission promised an action plan to promote EPL evaluation, revised its evaluation quality standards, and emphasized the essential role of ex-post evaluation (European Commission, 2007: 3; Fitzpatrick, 2012: 478). Yet, the European Court of Auditors (2010: 42) stated that 'only 24% of *ex-post* evaluations addressed issues related to the review of existing legislation'. Given the fact that legislation is the main policy instrument used by the European Union (Lodge, 2008: 282), this suggests that legislation remains an underemphasized component of the Commission's evaluation system.

The Commission restated its ambitions for ex-post evaluation, including EPL evaluation, several times (European Commission, 2010a; 2015: 255). In his political guidelines for 2010-2014, president Barroso (2009: 29) stressed that the Commission should 'match this huge investment in ex-ante assessment with an equivalent effort in ex-post evaluation'. Partly to this end, in 2009 co-ordination of the evaluation system shifted to the Secretariat-General (SG). Commission policy documents have come to picture the regulatory process as a 'cycle' consisting of the following consecutive stages: inception; design; legislation; implementation; and review; which should inform a new cycle (European Commission, 2013: 13; Smismans 2015: 11). The Commission's current strategy unfolds along the twin lines of evaluations of individual legislation and 'fitness checks' - evaluations of policy areas (European Commission, 2015: 254).

3. Theoretical expectations

Despite their theoretical potential, the systematic production of high-quality EPL evaluations also presents the Commission with a risk. Since EPL evaluations may uncover critical problems in the actual working of legislation, they may function as a 'dagger in the back' (Vedung: 1997: 108) for the European Commission. Member states, interest groups or the European Parliament (EP) may use unfavourable evaluations to demand the Commission to change or repeal legislation. Although the Commission itself (2013: 8) recognizes legislative repeal as a possible consequence, this possibility seems at odds with the Commission's alleged strategy of continuous legislative expansion (Lodge, 2008: 286; Majone, 1999: 2; Majone, 2005: 146). According to Majone (2005: 39), the logic of EU integration is one of 'integration by stealth' - the Commission continuously striving to protect and promote the supranational interest.

Accordingly, Majone argues that the Commission is relatively indifferent to actual policy outcomes, which instead are more the by-product of actions undertaken to advance the integration process (Majone, 2005: 107). Accordingly, as argued by Majone (2005: 107), 'policy evaluation (... plays) a very limited role in the EU policy process'.

A similar conclusion can be reached when applying principal-agent theory to the EU (Tallberg, 2003: 5-6). In this logic, EPL evaluations can be seen as an oversight procedure for the Council to control the Commission (Pollack, 1997: 109). Accordingly, the Commission may wish to avoid high-quality evaluations in policy areas where it wishes to maintain its information advantage (Pollack, 1997: 126).

The Commission thus faces a dilemma: to capture the potential of systematic high-quality evaluation it introduces the risk of undermining hard-fought legislative compromises and reopening legislative dossiers. This dilemma may diminish the incentive to produce objective evaluations aimed at fact-finding (Bovens et al., 2008: 323). Similar trade-offs have been shown for EU impact assessments (Boswell, 2008: 485-86; Radaelli and Meuwese, 2010: 145), and evaluations more generally (Bovens et al., 2008: 320; Vedung, 1997: 111-113).

At the same time, the Commission can be seen as the principal of the member states. In its role of guardian of the treaties, the Commission is responsible for monitoring and enforcing EU legislation. Given its lack of inspection powers, the Commission could use EPL evaluations to assess policy implementation or administration at the national level (Stame, 2008).

Turning to the rationalistic perspective on evaluation (Bovens et al, 2008), the production of systematic high-quality evaluation is likely to be hampered by problems of a more technical nature. The assumption in this literature is that evaluation initiation and quality depend on evaluability: how easy or difficult a policy is to evaluate (Summa and Toulemonde, 2002: 408). In the EU context, evaluability is threatened by divergent implementation by the member states (Fitzpatrick, 2012: 480; Summa and Toulemonde, 2002: 408) - which is likely when EU legislation contains discretion (Franchino, 2007: 1), or is ambiguous (Mastenbroek, 2003: 376-377). Furthermore, evaluability is hampered by the absence of a natural moment to evaluate regulations or directives - in contrast to policy programmes, which have fixed life cycles (Fitzpatrick, 2012: 480-481).

In sum, we argue that systematic high-quality EPL evaluation by the Commission is not likely to materialize, an exception being evaluations aimed at establishing member state implementation. This expectation forms the rationale for a meta evaluation of EPL evaluations in the EU setting.

4. Analytical framework

Our meta evaluation comprises two criteria: systematic coverage and evaluation quality. The first evaluation criterion is *systematic coverage* (Stern, 2009: 71). According to the European Commission's own evaluation standards (European Commission, 2015: 253-7), all activities addressed to external parties must be periodically evaluated in proportion with the allocated resources and the expected impact. In reality, evaluation coverage is seldom complete, because resources for evaluations are likely to be focused (Radaelli and Meuwese, 2010: 146). In line with our general expectation, we expect to see patchy evaluation, with no clear trend over time, and several politically sensitive laws not being evaluated. Owing to the political costs of evaluation, we expect evaluation to be primarily obligatory in nature, necessitated by a legislative evaluation clause. Additionally, since directives are likely to be more prone to divergent implementation by the member states than regulations (Fitzpatrick, 2012: 480), we will also explore whether the Commission prioritizes this type of evaluation.

Our second evaluation criterion is *evaluation quality*. Political pressures on evaluators may reduce evaluation quality (Schwartz and Mayne, 2005: 2; Versluis et al., 2011: 223). The European Commission, in its role as an agent, is expected to hide information from its principals to maximize its autonomy (Tallberg, 2003: 19). Turning to the rationalistic perspective, 'technical' impediments to methodological quality in the EU are limited availability of data owing to a lack of inspection powers for the Commission, and problems inherent to the methodologically daunting task of establishing the working and outcomes of legislation in an EU of 28 member states (Fitzpatrick, 2012: 480).

Given the underlying idea that EPL evaluations are to inform subsequent revision and possibly even repeal of legislation, evaluation quality is, firstly, a matter of the type of evaluation conducted. If evaluations are to seriously gauge the extent to which 'a treatment works in terms of reducing a problem', they must incorporate the administration, compliance,

and outcomes of legislation (Coglianese, 2012: 14) and go beyond mere process evaluations focusing on transposition, implementation and enforcement of EU legislation.

A second quality variable (European Commission, 2007: 5; Forss and Carlsson, 1997: 488; Schwartz and Mayne, 2005: 1-2) is methodological in nature. If evaluations are to credibly inform learning and adjustment, the evaluator must produce findings with a great degree of plausibility, which implies standard methodological conditions (Forss and Carlsson, 1997; Schwartz and Mayne, 2005). Yet, in light of our general expectation, the methodological quality of EPL evaluations is likely to be limited.

The first methodological quality aspect is *well-defined scope*, implying clear evaluation objectives (Schwartz and Mayne, 2005: 6). The second aspect is *measurement validity*, holding that operationalization and scoring adequately reflect the researcher's concept of interest (Adcock and Collier, 2001: 529; Schwartz and Mayne, 2005: 6). The third aspect is *external validity*: the degree to which results can be generalized to the entire population. Fourth, *reliability* means that the results of research are not distorted by random errors (Babbie, 1986: 109; Schwartz and Mayne, 2005). Hence, it is important that an evaluation can always be replicated (Golafshani, 2003: 599). The fifth aspect is *robust methodology*, which implies justification of the choice of methods (Schwartz and Mayne, 2005: 6). The final aspect concerns the actual analysis; the findings of the evaluation should be impartial and based on the evidence gathered (Schwartz and Mayne, 2005: 6). A key strategy for ensuring this is *triangulation*: using different types of data.

A third quality variable is *evaluation process* quality. Here, the first aspect is *stakeholder involvement*. Stakeholders can provide practical information about problems and solutions and helps to disseminate results (European Commission, 2015: 280-1; Schwandt, 1990: 178; Rossi et al., 2004: 35-36). A second process aspect is *public availability* of the report. This allows for external scrutiny and use of the insights by stakeholders, which enhances an evaluation's impact on the regulatory cycle.

The final quality variable concerns *usefulness* (Rossi et al., 2004: 35): if evaluations are to play a role in the regulatory cycle, they must address those in power to revise or repeal legislation. Accordingly, an evaluation report should contain (a) a clear *executive summary*, and (b) *useful recommendations* (Forss and Carlsson, 1997: 495).

5. Methods and data

Data

The method of this article is a meta evaluation: 'a systematic review of evaluations to determine their processes and findings' (Cooksy and Caracelli, 2005: 2). Unlike Adelle et al. (2012: 401), who performed a meta-study of evaluation literature, this article focuses on EPL evaluation reports. We encountered two obstacles to this research strategy. First, the EU does not have a fully operational database of evaluation reports (Smismans, 2015: 13). The official sources available mostly contain studies that are evaluations by the Commission's standards,[4] but lack broader studies with evaluative aspects following our more-encompassing definition. Second, assessing EPL evaluation coverage is complicated given the EU's vast legislative output. Therefore, we developed two novel datasets.

First, we constructed a dataset containing EPL evaluation reports. This dataset contains 216 EPL evaluations commissioned or conducted by the European Commission from 2000 to 2012. Evaluations merely containing prescriptions for foreign actors or the EU institutions themselves were excluded.[5] Six evaluations only available in French were left out so as to prevent bias owing to varying degrees of language abilities. Evaluation reports that only reported the results of other studies were discarded. Interim evaluations were incorporated unless an EPL evaluation by the same evaluator about the same law existed.

The evaluation reports were collected from the Commission's multi-annual evaluation overview (European Commission, 2010b), Commission annual evaluation overviews (European Commission, 2011), annexes of the Commission reports on the financial regulation (the so-called 318 report; European Commission 2012b), Commission work programmes,[6] and overviews on websites of the Directorates-General (DGs). We complemented this information with a dataset compiled by Eureval, a private evaluation company.[7] To maximize coverage, we conducted Google searches for evaluation reports of all directives and regulations adopted between 1996 and 2004[8]. Finally, we used the evaluation search engine from the Commission[9] and the online EU bookshop,[10] the latter of which yielded no additional studies. However, we did identify various new studies when searching for background documents underpinning

legislation adopted.[11] We double-checked our data-gathering method with the SG of the European Commission, which did not reveal any omissions.

To assess evaluation coverage, we constructed a second dataset containing a manageable set of important legislation adopted from 2000 through 2002.[12] Given the assumption that evaluation spending should be proportionate to the legislation concerned (Stern, 2009: 71), we excluded all delegated acts, because they generally are less important (Franchino, 2007: 80), as well as rectifications, amendments and secondary Council legislation.[13] We also excluded legislation addressing government institutions and non-EU members. The resulting dataset consists of 156 pieces of major directives and regulations.

Operationalization

Table 1 summarizes the operationalization. Concerning *systematic coverage*, we appended the EPL evaluations found to the dataset containing major legislation to establish the *proportion* of legislation evaluated. It should be stressed that one EPL evaluation report may cover multiple pieces of legislation. Additionally, we charted the *number* of evaluations per year and assessed how many evaluations were *obligatory* in nature, owing to the presence of an evaluation clause. We then established the *object* of evaluation, distinguishing between evaluations of regulations, of directives and of treaty articles.

Turning to the quality aspects, we analyse the *type of evaluation* by distinguishing between process and product evaluations (Vedung, 1997: 137, 165). Process evaluations either involve what Coglianese (2012: 14) calls regulatory administration or behavioural compliance. They may, in the EU context, concern transposition of directives into national law, operationalization by implementing authorities, actual application of the rules to specific cases, and/or enforcement by (sub)national authorities. Studies of behavioural compliance concern the extent to which the behaviour of regulated entities indeed is in line with the new standards. Product evaluations, on the other hand, focus on the actual outcomes of legislation. This type of evaluation may focus on four different types of *evaluands*: goal achievement; effectiveness; efficiency; and/or (side-)effects (Vedung, 1997: 54-55; 96). All evaluations were categorized as one of these nine types, based on the evaluation goal mentioned in the report.

Table 1: Operationalization of meta evaluation aspects

Variable	Aspect	Indicator	Source
Systematic coverage	Proportion	Proportion of major legislation 2000-2002 covered by EPL evaluations	Dataset 1
	Number	Number of EPL evaluations per year, 2000-2011	Dataset 2
	Obligatory	Proportion of obligatory evaluations	Evaluation clause in evaluated legislation
	Object	Regulation, directive or treaty article	Evaluation report
Type of evaluation	Process	Transposition, realization, implementation, compliance, enforcement, as reflected by central goal/question	Evaluation report
	Product	Achievement, effectiveness, efficiency, effects, as reflected by central goal/question	Evaluation report
Methodological quality	Well-defined scope	Clear problem definition: 1 point	Evaluation report
	Measurement validity	Operationalization into clear indicators present before presentation of the results: 1 point	
	External validity	Justified selection of member states: 1 point	
		Representative selection of cases within member states: 1 point	
		Less than 50% non-response at surveys or interviews: 1 point	
	Reliability	Replicability: 1 point	
	Robust methodology	Justification of methods: 1 point	
	Substantiated findings	Triangulation: 1 point	
	Aggregate quality	Aggregate score on the eight dimensions	
Process quality	Stakeholder involvement	No stakeholder involvement, stakeholder involvement for empirical information or stakeholder involvement for evaluation process as stated by the evaluation report	Evaluation report
	Availability	Evaluation report made available by SG, DG, or elsewhere	European Commission SG and DG websites
Usefulness	Clear executive summary	Executive summary of no more than 10 pages or report itself is less than 10 pages: 1 point	Evaluation report
	Useful recommendations	Possible actions within the power of the Commission: 1 point	

Given the quantitative nature of this article, operationalization of *methodological quality* was driven by the need for indicators that can be efficiently applied to a large number of cases, using the reports.[14] The problem in doing so is that evaluation quality can be compromised in many ways that are hard to observe without in-depth knowledge of the legislation evaluated. For example, selection of respondents may be biased, problem definitions restricted or analysis subjective. Because such biases are hard to identify in a quantitative fashion, we settled for formalistic indicators, which do not form a full guarantee against subjective evaluation. This choice is based on the principle that at least all methodological choices must be clear and transparent, so as to allow for replicability by third parties.

Concerning *well-defined scope*, we assessed whether the evaluation was guided by a clear problem definition. To assess *measurement validity,* we established whether the evaluation used a clear operationalization, i.e., stating clearly on what empirical observations its findings will be based. We used three indicators of *external validity*: arguments for selecting member states; arguments for selecting cases within member states; and a response rate for interviews or surveys above 50%.[15]

To measure *reliability*, we assessed whether an evaluation report publicly indicates its sources. All questionnaires, interview guides and lists of respondents or their organizations had to be attached to the report. As for *robust methodology*, we required the evaluation to explain *why* it used certain evaluation techniques. A simple explanation of the goal of the methods used was considered sufficient. To assess *substantiated findings*, we established whether evaluations used at least two of the following types of sources: written material; observations; surveys; and interviews. In cases where this was impossible, such as in the case of macroeconomic analysis, two different types of written sources were also considered sufficient. Consultation rounds with stakeholders were viewed as a type of interviews in case they were aimed at producing empirical information.

Turning to process quality, *stakeholder involvement* was measured by searching the text of each report for the terms 'stakeholder', as well as the most common forms of stakeholder involvement during evaluations.[16] We distinguished between three situations: (1) no stakeholder involvement; (2) stakeholder involvement in data provision; and (3) stakeholder

involvement in other phases of the evaluation process, such as methodological design or formulation of recommendations. Regarding *availability*, we identified were a report was found: on the website of the SG; a DG; or elsewhere.

Concerning usefulness, finally, we judged the *clarity of the executive summary* by assessing whether a summary of no more than 10 pages (University of South California, n.d.) was included in or attached to the report (Poptcheva, 2013: 4).[17] Reports with less than 10 pages of main text scored automatically on this condition. Concerning *useful recommendations*, we assessed whether the conclusion of each report suggested specific actions to be taken by the Commission in response to the findings of the evaluation.

Reliability

To improve the reliability of our results, 20 reports were double-coded using a codebook. Critical comparison of our scores did not yield systematic differences in scoring approach. This was substantiated by the fact that all kappa values in our subsequent intercoder reliability calculation were significant and higher than 0.4 (Neuendorf, 2002: 143) Therefore, we concluded that there are no significant problems with the reliability of our scoring exercise.

6. Analysis

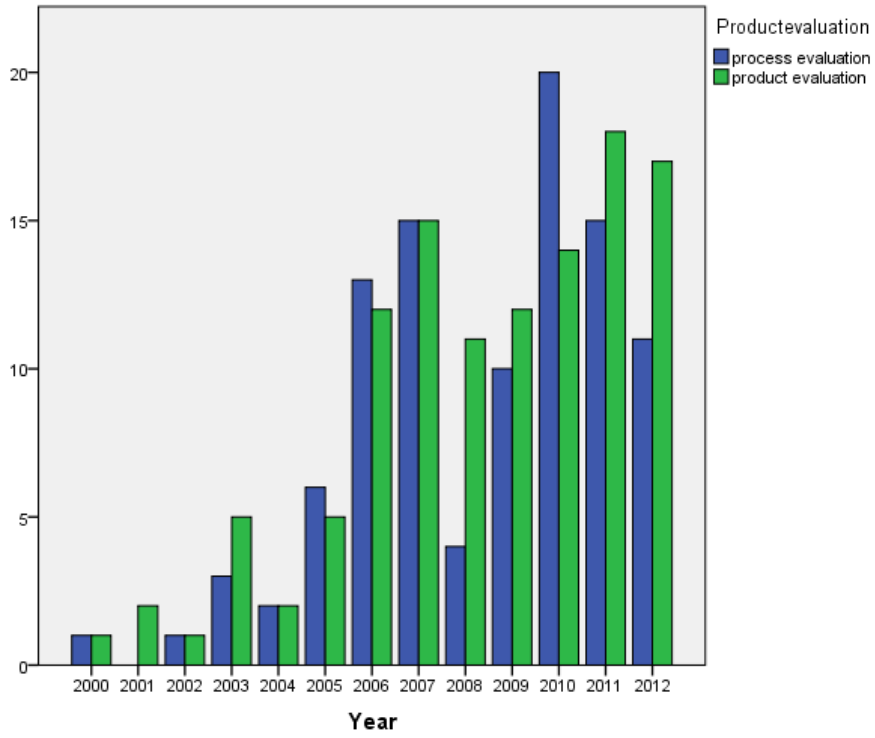
Coverage

Our data allow for systematic insights into how much important EU legislation was actually evaluated ex-post. Out of the 156 important pieces of legislation adopted between 2000 and 2002, 44 were evaluated once, while 8 were evaluated twice. This means that 33% of important EU legislation has been covered by an EPL evaluation. Accordingly, almost seven out of ten important legal rules have not been evaluated.[18] This figure is close to self-reported data of the Commission, according to which 29% of EU regulations have been evaluated ex-post (European Commission, 2013: 3). Amongst the 104 pieces of legislation that have not been evaluated, we find some politically sensitive pieces of law, such as the directives on facilitation of illegal immigration (2002/90) and on privacy and electronic communications (2002/58).

The evaluations conducted were not distributed equally over the years. As shown in Figure 1, relatively large numbers of evaluations were conducted in 2007 (30), 2010 (34), 2011 (33) and 2012 (28). For all other years, 25 evaluations or less were identified. The number of evaluations in the years 2000-2005 was relatively small (13.4%), which may be indicative of the increasing attention by the Commission to EPL evaluation from 2006 onwards. However, the increase has not been consistent- neither the 2007 Grybauskaité report, nor the move of co-ordination responsibilities to the SG or the attention provided to EPL evaluation by Barroso (2009: 29), seem to have affected output.[19] The proportion of product versus process evaluations varies somewhat from year to year, but the data do not show any clear trend towards one type of evaluation over time.

Turning to the *origins* of our evaluation reports, the large majority (81%) of the evaluations were based upon an evaluation clause. This suggests that legal obligation is the main motive to evaluate, and that the Commission's own initiative is a limited driver, in line with our expectations.

Figure 1: Number of evaluations per year, 2000-2012



As to the *object* of study, the majority of evaluations (57%) concerned directives, as expected. Evaluations of regulations were less frequent (42%); only one evaluation (report on EU citizenship 2012) concerned a treaty article. Some 21% of the 216 reports in our sample are product evaluations, some 31% of the evaluations containing both product and process elements. The remaining 48% of the studies are process evaluations, covering transposition, realization, implementation, compliance, and/or enforcement. So, although product evaluations constitute a larger share than expected, which indicates that the Commission goes beyond the production of mere 'enhanced implementation reports' (Fitzpatrick, 2012: 480), process evaluation overall seems more important than product evaluation, which is in line with our expectations.

Methodological quality

Table 2 shows the variation in methodological quality. The average score on quality is 4.1 on an eight-point scale. Some 43% of the reports received a score of 5 or higher, which could be considered 'sufficient' in the limited sense of an indication of overall quality. The average score was identical for process and product evaluations. However, communications to the EP and the Council differed in quality from other evaluations: the first group scored an average of 3.41 for methodological quality, while the second group scored an average of 4.46. As shown in the table, two cases scored no points at all. Both reports[20] indicate that the Commission considered the evaluation fairly pointless- either because a decision to repeal it had already been taken or because the law had been evaluated often before- but performed the evaluation nevertheless because of a legal obligation. The two reports scoring an 8[21] were both external studies of average length (60 pages), applying a combination of a large number of in-depth interviews with analysis of documentation provided by the member states.

To put these findings in (rough) comparative perspective, how do our general scores on quality compare to other studies of evaluation quality? Klein Haarhuis and Niemeijer (2009: 403) found that 79% of Dutch legislative evaluations were of sufficient quality, which is a more positive result. Studies about impact assessments are generally more negative and show that in some time periods over half are insufficient (Cecot et al., 2008: 414; Lee and Kirkpatrick, 2004: 17-19; Renda, 2006), which is closer to our result.

Table 2: Quality of evaluation reports

Score	Frequency	Percentage
0	2	1
1	7	3
2	25	12
3	42	19
4	47	22
5	57	26
6	21	10
7	13	6
8	2	1
Total	216	100

Table 3 summarizes the scores of the evaluation reports regarding the quality and usefulness conditions. Most reports scored well on justified selection of member states (73%), problem definition (67%), and triangulation (65%). Triangulation usually took the form of a mix of content analysis and either interviews or surveys. Some 62% of the reports had a clear operationalization. Representative case selection and response rate gained a positive score in 50% and 51% of the cases, respectively. Replicability (29%) and robust methodology (15%) were most problematic. Although most reports mentioned methods, motivation of the choice of methods was rare. When it came to replicability, few reports provided all the information needed to repeat the study- more often than not, either questionnaires or interview guides were absent.

Process quality

Some 39% of the 216 reports showed no sign of *stakeholder involvement*. In 51% of the cases, stakeholders provided information for the empirical part of the evaluation, but did nothing else. Only 9% of the reports showed that stakeholders were involved more deeply in the evaluation process, providing feedback on various aspects of the study. In sum, about 60% of the evaluations reported some kind of stakeholder involvement, without a large difference between process (60%) and product evaluations (61%).

Availability was not full for all evaluations. Some 44% of the evaluation reports could be found at the website of the Secretariat General, either through annual overview documents or

the search engine.[22] Out of the remaining 120 studies, 14 (7%) were available on general evaluation pages of individual DGs, 80 (37%) were only available on DGs' websites, such as web pages about specific policies, and 26 (12%) were only available elsewhere online. This means that in many cases reports are not communicated to the broadest possible audience. Out of all DGs with more than two reports in the dataset, DG Enterprise and Industry (ENTR) had the highest percentage of its evaluations listed in the Commission's search engine or centralized documents (13 out of 18 reports, or 72%), while DG Employment, Social Affairs & Inclusion (EMPL) scored the lowest in this regard (3 out of 15 reports, or 20%)

Usefulness

As shown in Table 3, the majority of the reports (76%) contained useful recommendations. For instance, the evaluation of regulation 2679/98 (strawberry regulation) clearly recommends the Commission to choose one out of four policy options, both in its main text and in a brief summary. Conversely, the evaluative study of directive 1997/81 on part-time work provides neither a summary nor final recommendations, despite its length of 276 pages. Accessibility of the reports scored lower, with 64% of the reports having a concise executive summary.

Table 3: Scores on methodological quality and usefulness

Aspect	**Number of reports that scored positively (%)**
Well-defined scope (clear problem definition)	145 (67)
Measurement validity (operationalization)	133 (62)
External validity 1 (justified selection of member states)	158 (73)
External validity 2 (representative case selection)	108 (50)
External validity 3 (response rate)	111 (51)
Substantiated findings (triangulation)	140 (65)
Reliability (replicability)	62 (29)
Robust methodology (justification of methods)	32 (15)
Usefulness (clear executive summary)	139 (64)
Usefulness (useful recommendations)	165 (76)

7. Conclusion

On paper, the European Commission attaches great importance to ex-post legislative evaluation, owing to its potential for the EU regulatory cycle in terms of legislative review and enforcement. At the same time, theory on EU governance informs us that EPL evaluations may present the Commission with a dilemma, given its strategy of continuous legislative expansion and problems of evaluability in the EU's multilevel setting. This article, therefore, has addressed the question whether the European Commission systematically produces high-quality EPL evaluations. To this end, it has reported on a meta evaluation of EPL evaluation reports against the background of a dataset of important EU legislation, producing the following insights.

First and foremost, coverage of EPL evaluations is patchy indeed: a mere 33% of major legislation adopted from 2000 to 2002 has been evaluated ex-post, despite the Commission's repeated pledges to evaluate all important legislation. Although the number of evaluations has shown an upward trend, this trend has not been clear and consistent, and seems unrelated to attempts of the Commission to enhance evaluation. Evaluation primarily seems a matter of obligation, given the importance of evaluation clauses. Although product evaluations constitute a larger share of reports than expected, process evaluation overall seems more important than product evaluation, which is in line with our expectations.

The methodological quality of the legislative evaluations was rather disappointing: only 43% of the evaluations scored sufficiently on a scale of 8 fairly conservative methodological conditions. Whereas most reports justified country selection and contained a clear problem definition, the majority of studies lacked replicability and robust methodology, which makes it hard to assess the objectiveness of results. Third, stakeholder involvement was completely lacking in 39% of the cases, with only 9% of the reports indicating real involvement of stakeholders in the evaluation process. Additionally, transparency of evaluation results was limited: only 44% of all EPL evaluations identified were available in the Commission's centralized search engine and annual overviews. Many reports thus are not communicated to the broadest possible audience. Usefulness of evaluations, finally, scored better than methodological quality, the majority of reports containing useful recommendations and having a concise executive summary.

In sum, it seems fair to conclude that the European Commission indeed has not yet lived up to its promise to close the regulatory cycle through EPL evaluation. A main recommendation to the Commission would be to enhance the coverage of evaluations, so that indeed all major regulations and directives are evaluated retrospectively. Methodological improvements of reports are called for as well, so that the findings of evaluations can withstand external scrutiny and credibly underpin future problem-solving. There also is great scope for enhancing external scrutiny by involving stakeholders in evaluation processes. To show the results of these investments to the public, finally, greater transparency is required, which could be achieved by including all EPL evaluations in the Commission's centralized search engine.

Furthermore, this research raises three crucial avenues for further research. First, the question of transparency begs closer scrutiny: what are the reasons for not fully publishing EPL reports? Is there a relationship with sensitivity of findings? Second, the concept of quality could be deepened by going beyond the formalistic quality conditions covered in our study. We propose to study a smaller number of EPL reports in depth, detailing qualitatively the sources of evaluation bias in EU legislative evaluations. Third, our results raise a crucial explanatory question: how can the observed variance in the initiation and quality of the ex-post legislative evaluations between pieces of legislation be explained? We intend to carry out follow-up research to establish the relative weight of political and more technical factors in producing variance in evaluation initiation and quality.

As pointed out throughout this article, principal-agent theory is a first useful framework to analyse the political perspective on evaluation. EPL evaluation can be a tool for the Commission to enforce its policies towards individual member states, but evaluation can also serve as a 'dagger in the back' when results are negative. Therefore, we would expect EPL initiation and quality to be a function of both the chances of non-compliance by the member states and the chances of policy reversal. Accordingly, we would expect initiation and quality of evaluations to be lower in policy areas where the EP is involved and the Council votes through unanimity, as these conditions make it harder to amend or repeal legislation.

Alternatively, initiation and quality may be affected by technical hurdles. In this perspective, legislative ambiguity and complexity, involvement of various implementers, as well as lack of resources, could combine to reduce the number and quality of evaluations. For

example, we would expect the initiation and quality of EPL evaluations to be higher if an impact assessment about the same topic was conducted before, as impact assessments often provide useful information for EPL evaluations (Smismans, 2015: 13). In addition, evaluation coverage may vary owing to differences in evaluation capacity of the Commission's DGs, which bear the main responsibility for EPL evaluation. The question is how demand for evaluation and supply for evaluation, in the sense of human capital and evaluation technology, interact at the various DGs and affect cross-DG variance in evaluation initiation and quality.

Taken together, these follow-up questions constitute a new research agenda on this so far neglected stage of the EU regulatory cycle.

Notes

[1] This article uses a broad definition of ex-post evaluation, comprising the actual outcomes of EU legislation, regulatory administration and compliance by regulated entities. This definition is broader than the one used by the European Commission (2007: 20).
[2] Exceptions are Fitzpatrick (2012), Summa and Toulemonde (2002), Toulemonde et al., (2005), Stame (2008), Stern (2009), and Poptcheva (2013).
[3] Although both EPL evaluations and programme evaluations are subject to the SG's evaluation guidelines (European Commission, 2015: 252-98), they differ in three respects: 1) While programme evaluations are often a shared responsibility between the Commission and the member states, the responsibility for legislative evaluation lies entirely with the Commission (Stern, 2009: 69); 2) While for programmes the moment to conduct an ex-post evaluation is fixed - usually at six years - EPL evaluations allow for more variation (European Commission, 2015: 256; Stern, 2009: 70-71); and 3)EPL evaluations are generally considered to be methodologically more difficult (Fitzpatrick, 2012: 481).
[4] See note 1 above.
[5] We excluded evaluations of legislation on reporting requirements for European institutions, on communication systems, and on the creation of agencies or research institutions.
[6] http://ec.europa.eu/atwork/key-documents/index_en.htm
[7] http://www.eureval.org
[8] We searched for the following key words: report on legislation; report on implementation legislation; evaluation legislation; evaluation implementation legislation; review legislation; review implementation legislation (replace the word 'legislation' with the name of the appropriate regulation, directive or decision).
[9]http://ec.europa.eu/smart-regulation/evaluation/search/search.do
[10] https://bookshop.europa.eu/en/home/
[11] The exact search method in the old version of Eur-lex was simple search→ prepatory acts→ search for the number of the legislation as keywords (e.g. 2004/82)→ refine search→ choose the European Commission as author.
[12] These years were chosen so as to maximize the chances of an evaluation having taken place. One should allow for transposition, in case of directives, as well as an average period of some five years before ex-post evaluation makes sense. Since the study was conducted in 2012, 2002 seemed a safe endpoint for the dataset.
[13] This refers to all legislation having its legal basis not in one of the EU-treaties, but in a regulation or directive.

[14] It must be noted that some information may be in the terms of reference instead of the report. Given the fact that terms of reference are not publicly available, we resorted to using the reports.
[15] This is in the middle of the percentages suggested in the literature, which vary between 20% (Valentine, 2009: 135) and 80% (Fowler, 2009: 51).
[16] The keywords used were 'stakeholder', 'consult*' and 'focus gr*'
[17] Although the Commission's guidelines (2004: 60) prescribe five pages maximum, we used a higher number, because the standards were adopted four years after our dataset's starting point.
[18] It must be noted that it is not unthinkable that these evaluations follow later. However, given our conservative timeframe, this is not very likely.
[19] Because we only entered the most recent version of evaluations that appear periodically (17 cases in the dataset), it is also possible that the numbers are slightly biased towards the more recent years.
[20] [COM(2007)253] and [COM(2007)287].
[21] Evaluations of the Measures under Regulation (EC) No. 951/97 and of the Data Protection Directive (95/46/EC).
[22] One explanation for this is the more restricted definition of evaluations used by the Commission. At the same time, 60% of the labelled as product studies we evaluations were also absent in the SG overviews.

References

Adcock R and Collier D (2001) Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95(3): 529-546.

Adelle C, Jordan A and Turnpenney J (2012) Proceeding in parallel or drifting apart? A systematic review of policy appraisal research and practices. *Environment and Planning C: Government and Policy* 30(3): 401-415.

Babbie E (1986) *The practice of social research (4th edition)*. Belmont, CA: Wadsworth.

Bachtler J and Wren C (2006) The evaluation of EU Cohesion Policy: Research questions and policy challenges. *Special Issue of Regional Studies* 40(2): 143-153.

Barroso JM (2009) *Political guidelines for the next Commission*. Available at: http://ec.europa.eu/commission_2010-2014/president/pdf/press_20090903_en.pdf (Accessed 15 June 2015).

Boswell C (2008) The political functions of expert knowledge: Knowledge and legitimisation in European Union immigration policy. *Journal of European Public Policy* 15(4): 471-488.

Bovens M, 't Hart P and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford handbook of public policy*. Oxford: University Press, pp. 320-335.

Cecot C, Hahn RW, Renda A and Schrefler L (2008) An evaluation of the quality of impact assessment in the European Union with lessons for the US and the EU. *Regulation and Governance* 2(4): 405-424.

Coglianese C (2012) *Evaluating the performance of regulation and regulatory policy*. Report to the Organization of Economic Cooperation and Development.

Cooksy LJ and Caracelli VJ (2005) Quality, context and use. Issues in achieving the goals of meta evaluation. *American Journal of Evaluation* 26(1): 31-42.

European Commission (2000) *Focus on results: Strengthening evaluation of Commission activities [SEC(2000)1051]*. Brussels: European Commission.

European Commission (2001) *European governance: A white paper [COM(2001)428]*. Brussels: European Commission.

European Commission (2002) *European governance: Better lawmaking [COM(2002)275]*. Brussels: European Commission.

European Commission (2004) *Evaluating EU activities: A practical guide for the Commission services*. Brussels: European Commission.

European Commission (2007) *Communication from the Commission from ms Grybauskaité in agreement with the president. Responding to Strategic Needs: Reinforcing the use of evaluation [SEC(2007)213]*. Brussels: European Commission.

European Commission (2010a) *Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Smart Regulation in the European Union [COM(2010)543 final]*. Brussels: European Commission.

European Commission (2010b) *Multi-annual overview (2002-2009) of evaluations and impact assessments*. Available at: http://ec.europa.eu/smart-regulation/evaluation/docs/multiannual_overview_en.pdf (Accessed 10 July 2015).

European Commission (2011) *List of evaluations 2010-2011*. Available at: http://ec.europa.eu/smart-regulation/evaluation/index_en.htm (Accessed 22 April 2012).

European Commission (2012a) *EU regulatory fitness [COM(2012)746 final]*. Brussels: European Commission.

European Commission (2012b) *Commission staff working document accompanying the document report from the Commission to the European Parliament and the Council on the evaluation of the Union's finances based on the results achieved [SWD(2012)383 final]*. Brussels: European Commission.

European Commission (2013) *Regulatory Fitness and Performance (REFIT): Results and next steps [COM(2013)685 final]*. Brussels: European Commission.

European Commission (2014) *Regulatory Fitness and Performance Programme (REFIT): State of Play and Outlook [COM(2014)368 final]*. Brussels: European Commission.

European Commission (2015) *Better Regulation Toolbox [SWD(2015)111 final]*. Brussels: European Commission.

European Court of Auditors (2010) *Impact assessments in the EU institutions: do they support decision-making? Special report no 3*. Luxembourg: European Court of Auditors.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

Forss K and Carlsson J (1997) The quest for quality - or can evaluation findings be trusted? *Evaluation* 3(4): 481-501.

Fowler FJ Jr. (2009) *Survey research methods (4th edition)*. Newbury Park, CA: Sage.

Franchino F (2007) *The powers of the Union. Delegation in the EU*. Cambridge: University Press.

Golafshani N (2003) Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report Volume* 8(4): 597-607.

Klein Haarhuis CM and Niemeijer E (2009) Synthesizing legislative evaluations. Putting the pieces together. *Evaluation* 15(4): 403-425.

Lee N and Kirkpatrick C (2004) A Pilot Study of the Quality of European Commission Extended Impact Assessments. *IARC Working Paper Series, Paper 8*. Manchester: Impact Assessment Research Centre.

Lodge M (2008) Regulation, the Regulatory State and European Politics. *West European Politics* 31(1-2): 280-301.

Majone G (1999) The regulatory state and its legitimacy problems. *West European Politics* 22(1): 1-24.

Majone G (2005) *Dilemmas of European integration: The ambiguities and pitfalls of integration by stealth*. Oxford: University Press.

Mastenbroek E (2003) Surviving the deadline: The transposition of EU directives in the Netherlands. *European Union Politics* 4(4): 371-396.

Neuendorf K (2002) *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Pollack MA (1997) Delegation, agency, and agenda setting in the European Community. *International Organization* 51(1): 99-134.

Poptcheva EM (2013) *Library Briefing. Policy and legislative evaluation in the EU*. Brussels: European Parliament.

Radaelli CM (2009) Rationality, power, management and symbols: four images of regulatory impact assessment. *Scandinavian Political Studies* 33(2): 164-188.

Radaelli CM and Meuwese ACM (2009) Better regulation in Europe: Between public management and regulatory reform. *Public Administration* 87(3): 639-654.

Radaelli CM and Meuwese ACM (2010) Hard questions, hard solutions: Proceduralisation through impact assessment in the EU. *West European Politics* 33(1): 136-153.

Renda A (2006) *Impact assessment in the EU: The state of the art and the art of the state*. Brussels: Centre for European Policy Studies.

Rossi PH, Lipsy MW and Freeman HE (2004) *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Schwandt TA (1990) Defining "quality" in evaluation. *Evaluation and Programme Planning* 13(2): 177-188.

Schwartz R and Mayne J (2005) Assuring the quality of evaluative information: theory and practice. *Evaluation and Programme Planning* 28(1): 1-14.

Smismans S (2015) Policy Evaluation in the EU: The Challenges of Linking Ex Ante and Ex Post Appraisal. *European Journal of Risk Regulation* 6(1): 6-14.

Stame N (2008) The European project, federalism and evaluation. *Evaluation* 14(2): 117-140.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 67-85.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation*. New Brunswick: Transaction, pp. 407-424.

Tallberg J (2003) *European governance and supranational institutions: Making states comply*. Abingdon, Oxon: Routledge.

Toulemonde J, Summa-Pollitt H and Usher N (2005) Triple check for top quality or triple burden? Assessing EU evaluations. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction, pp. 69-90.

University of South California (n.d.) *Executive Summary*. Available at: http://libguides.usc.edu/content.php?pid=83009&sid=1481087) (Accessed 10 July 2014).

Valentine JC (2009) Judging the quality of primary research. In: Cooper H and Hedges LV (eds) *The handbook of research synthesis (2nd edition)*. New York: Russel Sage Foundation, pp. 130-140.

Vedung E (1997) *Public policy and evaluation*. New Brunswick: Transaction.

Versluis E, Van Keulen M and Stephenson P (2011) *Analyzing the European Union Policy Process*. Houndmills: Palgrave MacMillan.

Chapter 3: Evaluation capacity in the European Commission

Stijn van Voorst

Published as: Van Voorst S (2017) Evaluation capacity in the European Commission. *Evaluation* 23(1): 24-41.

Abstract

Ex-post evaluations are a potential tool to improve regulatory interventions and to hold rule-makers accountable. For these reasons the European Commission has promised to systematically evaluate its legislation, but it remains unclear if actual evaluation capacity is being built up in the Commission's Directorates-General (DGs). This article describes and explains the variation in evaluation capacity between the DGs by applying a theoretical model of evaluation capacity developed by Nielsen et al. (2011) to the European context. To gain an in-depth understanding of the Directorates-General's evaluation capacity, 20 Commission officials were interviewed. The results show that there is much variation in the extent to which Directorates-General prioritize evaluation as well as in the amount of human and technological capital that they invest in evaluation. Further analysis using fuzzy-set Qualitative Comparative Analysis reveals that part of this variation can be explained by the Directorates-General's total budgets, suggesting that Directorates-General with a tradition of evaluating spending programmes also attach more importance to legislative evaluations.

1. Introduction

Ex-post evaluations are a potential tool for improving legislation, as they can be used to learn about the implementation and the actual impact of regulatory interventions (Fitzpatrick, 2012: 480). Evaluations can also help to hold regulators accountable for their actions (Summa and Toulemonde, 2002: 409) and to control those who implement legislation (Stern, 2009: 76). These purposes of evaluation are especially relevant for EU, which has limited access to financial

and communicative instruments and therefore relies heavily on legislative policies (Fitzpatrick, 2012: 489; Majone, 1999: 1).

Legislative evaluation is of particular importance to the European Commission, which bears the main responsibility for evaluation in the EU (Stern, 2009: 71). As an unelected body, the Commission has the constant need to show the added value of its policies (Scharpf, 1999: 187), for which ex-post evaluations can be a useful tool (Mastenbroek et al., 2016). Therefore, it is not surprising that the Commission has repeatedly stepped up its rhetoric in the field of ex-post legislative evaluations (Fitzpatrick, 2012: 489; Højlund, 2015: 40-4). In 2007 it pledged to evaluate not only its spending programmes, the evaluation of which has been common practice since the 1980s, but also non-spending activities like legislation (European Commission, 2007: 4; Højlund, 2015: 44). More recently the Commission (2013: 7; 2015: 17) has even promised to conduct evaluations of entire regulatory frameworks. However, existing research shows that these high ambitions may not always be realized. Available figures from the Commission (2013: 3) and academic research (Mastenbroek et al., 2016) indicate that the Commission has only evaluated about a third of the legislation that it should evaluate. Moreover, the quality of these evaluations seems to vary (Mastenbroek et al., 2016).

Theoretically, one explanation for variation in the initiation of evaluations lies in variation in evaluation capacity (Mastenbroek et al., 2016; Pattyn, 2014: 348). Evaluation capacity can be defined loosely as the presence of sufficient means and procedures for ensuring that evaluation and its appropriate uses are ordinary and ongoing (Stockdill et al., 2002: 14). It is a topic that has been studied in various settings, including non-profit organizations (e.g. Carman and Fredericks, 2010; Taylor-Ritzler et al., 2013), local and national governments (e.g. Bourgeois and Cousins, 2013; Nielsen et al., 2011) and international organizations (e.g. Taut, 2007). However, aside from a few sections in texts about the evaluation system of the EU (Stern, 2009: 71-2, 79-82; Summa and Toulemonde, 2002: 420-22; Toulemonde et al., 2005: 77-9) and evaluation use in the EU (Borrás and Højlund, 2015: 111; Technopolis, 2005: 45), there is little literature about the European Commission's capacity to evaluate, in particular when legislative evaluations are concerned.

Information provided by official sources is equally scarce. In 2007, the Commission reported that it had 140 full time equivalents (fte) working on evaluation, with a total budget of

€45 million (European Commission, 2007: 17), and that these numbers were gradually increasing because of investments made in trainings and networks (2007: 15-8). However, from this time onward the Commission has presented no more systematic data on its evaluation capacity, and it has never presented data on its capacity for legislative evaluations specifically.

This article seeks to fill this gap in our knowledge by answering the following questions: (1) to what extent do the Commission's Directorates-General (DGs) vary in their evaluation capacity? and (2) how can this potential variance be explained? The focus on DGs stems from the fact that they bear the main responsibility for conducting and outsourcing evaluations in the Commission (Stern, 2009: 71). The Commission's (2015: 268-88) guidelines for evaluation - which were first adopted in 2004 and updated by the Commission's Secretariat-General (SG) in 2015 - require each DG to maintain an evaluation function with sufficient financial and human capital. However, in the end it is up to the DGs how they fulfil these requirements (Stern, 2009: 71). Therefore, variation between the Commission's DGs is crucial for understanding the functioning of legislative evaluation in the EU.

Building on an existing model (Nielsen et al., 2011: 326-327), this article splits evaluation capacity between evaluation demand and evaluation supply. Both concepts are quantified on a scale of one to fifty points to allow for comparisons between DGs and to make the results useful for future research. Data were collected through in-depth interviews with twenty evaluation-related officials from seventeen DGs responsible for legislation. The results not only provide a complete overview for 2014, but also show how evaluation capacity in the DGs has developed since 2000. The findings show that there is much variation between DGs in both the way in which they organize their evaluation-related procedures and the means which they invest in evaluation. Further analysis using fuzzy-set Qualitative Comparative Analysis (fsQCA) reveals that part of this variation can be explained by differences in the budgets of DGs, suggesting that DGs with a strong tradition in the field of evaluating spending programmes also have more capacity for legislative evaluation.

2. Theoretical framework

Selection of a model

A commonly used definition of evaluation capacity is that it is 'a system of guided processes and practices for ensuring that evaluation and its appropriate uses are ordinary and ongoing' (Stockdill et al., 2002: 14). Beyond such general definitions, however, evaluation capacity is a very ambiguous concept, with frequent debates among authors about what its main components are and how they should be measured (Nielsen et al., 2011: 324; Taylor-Ritzler et al., 2013: 192). In this article, the model of Nielsen et al. (2011: 328) will be used to measure evaluation capacity, as it has four key advantages which are relevant in the context of the European Commission. Firstly, the model is meant to measure evaluation capacity at the organizational level (Nielsen et al., 2011: 326). Since the aim of this article is to measure variation among the DGs of the Commission - organizational entities with their own evaluation policies - this focus on organizational aspects makes the model suitable for the research question at hand. Secondly, the model of Nielsen et al. (2011: 330) was created in the context of public sector organizations. Many other models published in recent years (e.g. Bourgeois and Cousins, 2013; Taylor-Ritzler et al., 2013) focus on non-profit organizations in the US, which only conduct programme evaluations and are therefore hard to compare to the EU. Thirdly, the model allows evaluation capacity to be quantified on a scale of one to one hundred points, making it easy to compare capacity between organizations and making the results useful for future research. Fourthly, the validity of the indictors of the model of Nielsen et al. (2011: 334-7) was thoroughly tested using factor analysis.

Besides these four advantages, there are three potential drawbacks to Nielsen et al.'s model. Firstly, as the model is focused on organizational aspects, it ignores aspects of capacity related to single evaluations, such as the value that individual evaluation managers attach to learning (Bourgeois and Cousins, 2013: 299; Taylor-Ritzler et al., 2013: 192). However, because more than 200 ex-post legislative evaluations have been conducted in the EU between 2000 and 2012 alone (Mastenbroek et al., 2016), it would be impossible to measure indicators for every single evaluation. Secondly, the model is mostly focused on the minimum requirements that must be in place for evaluations to be embedded in an organization. Even when all these

requirements are met, there is no guarantee that this will result in sound evaluations being produced and put to use. This should be remembered when interpreting the results. Thirdly, the model was developed in the context of Danish local governments, meaning that it cannot be applied entirely to the EU level. In particular, the fact that most of the Commission's evaluations are outsourced (Stern, 2009: 69) had to be accounted for, resulting in some adaptions to the operationalization of the model which are further described below. However, most aspects of Nielsen et al.'s model could be applied to the Commission, as its evaluation staff is explicitly required to be able to conduct internal evaluations and to scrutinize external evaluators whenever needed (European Commission, 2015: 268, 288).

Describing the model

Following other authors (e.g. Boyle et al., 1999: 11), Nielsen et al. distinguish between evaluation demand and evaluation supply (2011: 327). Evaluation demand refers to the fact that an organization considers evaluations valuable, while evaluation supply refers to the presence of sufficient means to evaluate (Nielsen et al., 2011: 326-7; Summa and Toulemonde, 2002: 422-3). In this context, 'means' refer to the staff responsible for evaluation and its methodological tools. Supply-side and demand-side conditions are equally important in determining evaluation capacity, as evaluations only come into existence when both interact, and therefore both are awarded fifty points in the model (Nielsen et al., 2011: 330).

Evaluation demand consists of two dimensions: the extent to which an organization has the explicit aim to evaluate (**evaluation goals**) and the extent to which evaluation is embedded in the daily functioning of an organization (**structure and processes**) (Nielsen et al., 2011: 326-7). Both dimensions are considered equally important in determining demand and are therefore awarded exactly 25 points (Nielsen et al., 2011: 330).

Evaluation goals, the first dimension of evaluation demand, consists of three main aspects (Nielsen et al., 2011: 328). The first is the amount of *formalization*: do official documents describe if and when evaluations must be conducted? A second aspect is the *utilization* of results, as evaluations are ultimately meant to be used, at least officially. Finally, the number of *evaluation purposes* stated by an organization is relevant. After all, evaluations

can be used not only to create policies, but also to allocate resources, improve accountability and set priorities for the future (Nielsen et al., 2011: 180-84).

Structure and processes, the second dimension of evaluation demand, can be split into two aspects: the presence of an independent *evaluation unit* and the amount of *financial priority* that an organization attaches to evaluation (Nielsen et al., 2011: 330). Nielsen et al. also include the number of functions which an evaluator performs besides his core task in this dimension, but this aspect is left out in this article because in the EU most evaluations are conducted by external consultants.

Evaluation supply, the second condition of evaluation capacity, consists of two dimensions: the skills of those performing evaluations within an organization (**human capital**) and the non-human tools that allow evaluations to be performed (**evaluation technology**) (Nielsen et al., 2011: 327). Human capital (35 points) is more important than technology (15 points), as non-human tools are ultimately useless if there are no people who can apply them properly (Nielsen et al., 2011: 330).

When it comes to **human capital**, the first dimension of evaluation supply, three aspects are important: the *number of full-time employees* working on evaluations, the *evaluation trainings* completed by these employees and their *evaluation-related expertise* (Nielsen et al., 2011: 330). Nielsen et al. also include the formal education level of an organization's employees in this dimension, but this aspect is left out here as all the Commission's staff should have a master's degree, meaning there is little variation.

Evaluation technology, the second dimension of evaluation supply, consists of two aspects: the number of different *evaluation methods* (e.g. interviews, questionnaires) that are used and the application of any explicit *evaluation models* by an organization (Nielsen et al., 2011: 327; for examples of models, see Fitzpatrick, 2012: 481). Nielsen et al. also include the presence of evaluation software in this dimension, but since EU evaluations are usually outsourced this aspect was irrelevant for this study. Table 1 summarizes the model as it was used in this article.

Table 1: model of evaluation capacity

Condition	Dimension	Aspect
Evaluation demand	Evaluation goals (25p)	Evaluation purposes (8p)
		Formalization (7p)
		Utilization (10p)
	Structure and processes (25p)	Evaluation unit (9p)
		Financial priority (16p)
Evaluation supply	Human capital (35p)	Number of full-time employees (10p)
		Evaluation training (10p)
		Evaluation-related expertise (15p)
	Evaluation technology (15p)	Evaluation methods (10p)
		Evaluation models (5p)

Explaining evaluation capacity

Besides describing the variation in evaluation capacity between the Commission's DGs, this article also seeks to explain it. Although there is no single theoretical framework for this purpose, three separate explanations can be derived from the literature: the amount of legislation to be evaluated (functionalist logic), the presence of a tradition of evaluating spending programmes (historical institutionalism) and the sensitivity of the DG's policies (political rationality).

Firstly, following a functionalist logic, the *amount of legislation* that has to be evaluated by a DG could influence its evaluation capacity. Since evaluation is compulsory for most EU legislation (European Commission, 2015: 261), it can be expected that DGs responsible for more legislation will have more evaluation demand (hypothesis 1) and supply (hypothesis 2).

Secondly, the extent to which an organization has a *tradition of evaluating spending programmes* is a possible explanation for variation in evaluation capacity, since building such capacity is a long-term investment (Preskill and Boyle, 2008: 451). Because EU evaluations have their origins in the field of spending programmes, the DGs that spend most money have the longest experience with evaluation (Fitzpatrick, 2012: 479; Stern, 2009: 69). Therefore, it is expected that DGs with a stronger tradition of evaluating spending programmes also have more demand (hypothesis 3) and supply (hypothesis 4) for legislative evaluations. This argument follows the logic of historical institutionalism: the policies of an organization are bound by the decisions which it made in the past (Nugent, 2010: 438).

Thirdly, evaluation capacity could be influenced by the *political sensitivity of the DG's policy area*. The Commission is often assumed to follow a strategy of legislative expansion: because it lacks strong financial or communicative instruments, it tends to focus on expanding EU law to encourage European integration (Majone, 1999: 65). Evaluations are a potential threat to this strategy, as they can be used as an argument to roll back policies in case their findings are negative (Weiss, 1993: 94). This idea is closely linked to the political rationality of evaluations: evaluations are not neutral objects, but can be used to threaten or defend the interests of actors in the policy process (Bovens, 2008: 320; Nielsen et al., 2011: 94). Bureaucracies may try to avoid them when they threaten to undo the results of previous political investments or negotiations (Weiss, 1993: 95-6). Following this logic, it can be expected that DGs dealing with sensitive policies areas will have less evaluation demand (hypothesis 5) and supply (hypothesis 6).

3. Methods and data

Data collection

Empirical data were gathered through face-to-face interviews with the main coordinator of ex-post legislative evaluations in each DG (seventeen in total). Although it was attempted to also speak with the head of the evaluation function of each DG, this was only possible in three cases: most heads of unit referred back to their coordinator for legislative evaluation because the requested information was highly detailed. Interviews with three DGs could only be conducted by phone or e-mail[1]. The data provided by the respondents were always checked by using available documentation about the evaluation policies of the DGs (such as guidelines and annual activity reports). Such documents were usually found on the DGs' websites, but respondents were also asked to provide additional documents. In three cases, an indicator could be measured only through online documents[2].

Since this article focuses on legislative evaluations, only DGs responsible for major legislation were included. To find out which DGs meet this requirement, a self-constructed dataset of European regulations and directives from 2000-2014 was used (see chapter 2 of this dissertation for a detailed description). The dataset excludes amendments, rectifications,

implementing legislation, repeals, and legislation concerning individual countries, because such small acts are rarely evaluated (Stern, 2009: 71). Using the online database Eur-lex, each piece of legislation was linked to the DG which initiated it. Only DGs appearing at least once in this way were included in the research. DGs dealing with foreign affairs or the Commission's internal functioning were excluded, as their legislation is not aimed at citizens and therefore follows a different logic concerning evaluation. Applying these criteria, the study focusses on seventeen DGs existing in 2014: Agriculture (AGRI), Communications and Technology (CONNECT), Competition (COMP), Economic and Financial Affairs (ECFIN), Employment, Social Affairs and Inclusion (EMPL), Energy (ENER), Enterprise and Industry (ENTR), Environment (ENV), Eurostat (ESTAT), Home Affairs (HOME), Justice (JUST), Maritime Affairs (MARE), Internal Market (MARKT), Mobility and Transport (MOVE), Health and Consumers (SANCO), Taxation (TAXUD) and Trade (TRADE).

Operationalization

To operationalize the aspects of evaluation capacity described in the theoretical framework section, the indicators used by Nielsen et al. (2011: 330-1) and their relative weights were used as much as possible (Ramboll Management Consulting, 2011: 2). However, four adaptions were made to fit the model to the specific context of this study. Firstly, since the exact number of evaluations conducted by the Commission is unclear (Mastenbroek et al., 2016), all indicators requiring knowledge about numbers of evaluations were removed. Secondly, since evaluations in the EU are often outsourced (Stern, 2009: 71), two different indicators were used to measure evaluation-related expertise. Thirdly, the indicators for evaluation-related trainings and utilization were dichotomized because some collected data for these indicators were unspecific. Fourthly, the indicator for evaluation models was changed to specify what an 'evaluation model' means in the context of the EU. When an indicator was removed or added to the model, the number of points awarded to the other indicators inside the same dimension was increased or decreased proportionally.

The aspect of *formalization* was measured by asking if the DG has an official planning for future legislative evaluations (4 points) and any formal rule for when legislation should be evaluated when this is not compulsory (3 points). The aspect of *utilization* was measured by

asking if there is a standardized procedure for employees of the DG to respond to results of legislative evaluations (10 points). *Evaluation purposes* were measured by asking the respondents what aims legislative evaluation has in their DG. Improving policies, increasing accountability, efficient resource allocation, political supervision, long-term learning and setting priorities all qualify as different purposes (8 points).

The presence of an independent *evaluation unit* was measured by asking if the DG has a unit or subunit for which ex-post evaluation and related issues (like ex-ante evaluation) are its core task (9 points). Although each DG must have an evaluation function (European Commission, 2007: 16), this does not necessarily take the form of a specialized unit, so there is room for variation here. *Financial priority* was mapped by asking how much money each DG spends on an average ex-post evaluation of one regulation or directive (16 points). There are no hard standards for this kind of expenditure in the EU, so the DG with the highest expenditure per evaluation was used as a benchmark and other scores were adapted proportionally.

The *number of fte* working on evaluation was measured by asking how many people (in fte) work for the centralized evaluation function of each DG (10 points). Although other employees can also spend time on evaluations, their work is too fragmented to measure. The aspect of *evaluation training* was measured by asking if the DG organizes any evaluation trainings (10 points). *Evaluation-related expertise* was measured by asking in how many evaluation-related networks the DG's employees participate, as such networks are an important way of building expertise (Stern, 2009: 71) (9 points). To measure the external expertise available to each DG, the average number of external companies that bid for its legislative evaluation was asked for (6 points).

The aspect of *evaluation methods* was measured by checking if the DG has guidelines on how to conduct ex-post legislative evaluations in its policy field, with 2 points awarded per method described (10 points). The number of methods is relevant here because Commission officials must be able to scrutinize a broad range of external evaluations (European Commission, 2015: 288). The aspect of *evaluation models* was measured by checking if the DG has any written guidelines for modelling causal effects in legislative evaluations (5 points). Since all legislative evaluations of the Commission (2015: 53) are supposed to map causality, this indicator is relevant for each DG. For both evaluation methods and models, the number of

points awarded is halved if the DG only has guidelines for ex-post evaluation in general. The operationalization is summarized in Table 2.

Table 2: operationalization of evaluation capacity. Indicators highlighted with an asterisk were measured specifically for legislative evaluations; other indicators could only be measured for ex-post evaluation in general

Dimension	**Aspect of capacity**	**Indicator measured during interviews (max 100p)**
Evaluation goals (25p)	Evaluation purposes (8p)	Number of purposes of legislative evaluation stated (2p per purpose)*.
	Formalization (7p)	Presence of evaluation planning (4p, yes/no)*.
		Presence of guidelines about when to evaluate legislation (3p, yes/no)*.
	Utilization (10p)	Presence of a procedure for responding to legislative evaluation results (yes/no)*.
Structure and processes (25p)	Evaluation unit (9p)	Presence of a unit for which ex-post evaluation is a core task (yes/no).
	Financial priority (16p)	Money spent on an average evaluation of one regulation or directive as a % of the money spent by the DG with the highest expenditure*.
Human capital (35p)	Number of full-time employees (10p)	Number of fte working for centralized evaluation function (2p per fte).
	Evaluation training (10p)	Existence of a DG-level evaluation training (yes/no).
	Evaluation-related expertise (15p)	Number of evaluation-related networks (9p maximum, 3p per network).
		Number of external companies that bid for legislative evaluations of the DG (6p maximum, 1p per company)*.
Evaluation technology (15p)	Evaluation methods (10p)	Presence of internal guidelines on evaluation methods (10p, 2p per method)*.
	Evaluation models (5p)	Presence of internal guidelines on evaluation models present (yes/no)*.

As for the explanatory conditions, the *amount of legislation* was measured via the number of major regulations and directives initiated by each DG over the period 2000-2014. To measure this, the self-constructed dataset of legislation which was already described above was used. The extent to which a DG has *a tradition of evaluating spending programmes* was estimated through the size of its budget, which was retrieved from the annex of each DG's annual activity report[3]. The *political sensitivity of the DG's policy area* was measured by looking at the policy field which is handled by each DG. In the EU, some policy fields are (partly) dealt with through unanimity voting in the Council because they are related to the sovereignty of the member states, which means they are considered sensitive issues. This mostly concerns justice and home affairs, social policies and taxation, so the DGs dealing with these topics (JUST, HOME, EMPL and TAXUD) were coded as sensitive while the other DGs were coded as non-sensitive (Nugent, 2010: 308).

Method

Fuzzy-set Qualitative Comparative Analysis (fsQCA) was used to explain evaluation capacity[4], for two reasons. First, this method is useful for mapping the various combinations of causal conditions that can explain variation in a given outcome (in this article: evaluation demand and supply). Second, while the sample of seventeen DGs is too small to allow for regression analysis, it is large enough to make fsQCA a feasible method (Ragin, 2008: 9).

FsQCA allows for the use of continuous scales if they are transformed to vary between zero and one. Through the so-called direct method of transformation, conditions can be transformed if appropriate values are derived from the literature for the scores of zero (non-membership), 0.5 (cut-off point) and one (full membership) (Ragin, 2008: 85). Since the presence of political sensitivity is measured dichotomously in this article, the only conditions that require transformation are the presence of a large amount of legislation, the presence of a large budget, and the two outcomes (the presence of high evaluation demand and high evaluation supply).

DG Environment (ENV) is a prime example of a DG with a large amount of legislation. Because the aim of environmental policy is to regulate and prevent polluting behaviour from citizens and companies, the DG is responsible for a large number of regulations and directives in

the fields of waste, water quality, chemicals, noise, genetic manipulation and biodiversity (Nugent, 2010: 346-50). Therefore, the observed number of legal acts of DG Environment (76), the second highest in the data, was used as the score for full membership.

DG Agriculture (AGRI) is the classic example of a DG which relies heavily on spending. To support European farmers and make their activities sustainable, the DG uses a mix of direct payments, refund operations, rural development programmes and other subsidies. Together agricultural policies account for 40% of the EU budget. Therefore, the observed budget of DG AGRI (about €58 billion in 2013) represents full membership (Nugent, 2010: 353-63).

DG Home Affairs (HOME) is a prime example of a DG with a medium-sized amount of legislation and a medium-sized budget. The DG aims to halt crime and illegal migration, which requires numerous regulations about cooperation between member states - the Dublin regulation on migration being one example (50 observed legal acts in the data). However, the DG also manages financial activities such as the European refugee fund and the European return fund (observed budget: about €1 billion) (Nugent, 2010: 335-9). Therefore, the observed values of DG Home Affairs were used to represent the cut-off points for the explanatory conditions. This leaves four DGs which spend several billions in the group with 'high budgets' (the DGs for agriculture, technology, employment and enterprise), and a larger group of eleven DGs which spend 'just' a few hundred million euros in the group with 'low budgets'. The second group of DGs includes many DGs that rely on a large amount of legislation (70-90 acts) rather than high budgets for their policies (i.e. the DGs for environment, the internal market, health affairs and infrastructure).

There is also a group of seven DGs with neither a high budget nor a high amount of legislation. DG Competition is a prime example of this. Its main activities are decisions to allow or forbid state aid and company mergers, which requires only a handful of regulations (7) and a small budget (about €5.6 million in 2013) (Nugent, 2010: 327-8). Therefore, its observed values were used to represent non-membership for the explanatory conditions.

Evaluation demand and supply are transformed by dividing their scores by fifty, which can be done for such self-constructed scales (Kogut et al., 2004: 123). Because all the cut-off points presented in this section are to some extent arbitrary – there is no strong set-theoretical

knowledge for any of these conditions – alternative cut-off points will also be tested during the analysis to see how this affects the results.

4. Results

This section presents the empirical results for the indicators for evaluation capacity listed above. For DG ECFIN, data on all indicators measured specifically for legislative evaluations are missing, as this DG will only start evaluating its legislation in the near future. For DG ESTAT there is no data on the number of external companies bidding for legislative evaluations and the amount of money spent on an average evaluation, as all its legislation is evaluated internally and no consultant is paid. Table 3 and 4 summarize the results.

Concerning *evaluation purposes*, all DGs mentioned the improvement of policies and the need to be accountable to the Council and the European Parliament as important aims of legislative evaluation. About half of all DGs also mentioned setting political priorities as a purpose of legislative evaluation. Some DGs mentioned other aims as well, such as basic learning (COMP, TAXUD) and efficient resource allocation (AGRI).

As for *formalization*, all DGs have a planning for future legislative evaluations as a part of their annual management plans, as this practice is enforced by the SG. Some DGs publish this part of their annual management plan online, while others keep it internal. Seven DGs have guidelines stating after how many years a piece of legislation should be evaluated, which was between five and seven years in all cases. For other DGs an evaluation is generally initiated only when an evaluation clause makes this compulsory or when revision appears necessary.

Concerning *utilization*, nine DGs have an official follow up-procedure which applies to legislative evaluations. This usually takes the form of a requirement to write an action plan or fiche by the main policy unit involved in the evaluation, although in DG SANCO the plan is sometimes written by the evaluation unit and in DG EMPL it is a joint responsibility. Such actions plans usually require approval at the management level. For DGs without a follow-up procedure, legislative evaluations are often followed directly by an impact assessment (an obligatory ex-ante evaluation of legislative amendments) or by no action at all.

Table 3: results for evaluation demand (2014)

DG	Evaluation purposes	Planning	When evaluate	Utilization	Evaluation unit	Financial priority
AGRI	3	Yes	Yes	No	Yes (1998)	400.000
CONNECT	3	Yes	Yes	Yes	No	200.000
COMP	4	Yes	No	No	No	225.000
ECFIN	-	-	-	-	Yes (2005)	-
EMPL	2	Yes	No	Yes	Yes (1998)	300.000
ENER	2	Yes	No	Yes	No	250.000
ENTR	3	Yes	Yes	Yes	No	260.000
ENV	3	Yes	Yes	No	No	250.000
ESTAT	2	Yes	No	Yes	Yes (2005)	-
HOME	3	Yes	No	No	No	350.000
JUST	2	Yes	Yes	No	No	200.000
MARE	2	Yes	No	No	No	200.000
MARKT	3	Yes	Yes	Yes	Yes (2008)	240.000
MOVE	2	Yes	No	Yes	Yes (2007)	195.000
SANCO	2	Yes	Yes	Yes	No	200.000
TAXUD	2	Yes	No	No	Yes (2010)	210.000
TRADE	3	Yes	No	Yes	No	175.000

Of all the DGs examined, only DG AGRI has a *unit fully dedicated to ex-post evaluation*. DG ECFIN, EMPL, ESTAT, MARKT, MOVE and TRADE have units responsible for ex-post evaluation and related issues (like impact assessments), while for other DGs the evaluation function is located together with broad support functions like finances or strategy. DG AGRI also leads when it comes to *financial priority*, reporting the highest budget for an average evaluation of one regulation or directive (€400.000) because it often requires case studies in each member state. DG HOME also reports a high budget per evaluation (€350.000), while the other DGs vary between €175.000 and €300.000. This data should be interpreted with some caution, as budgets show much variation.

Table 4: results for evaluation supply (2014)

DG	Evaluation training	Fte	Number of networks	Number of bids	Method guidelines	Model guidelines
AGRI	No	15	3	5	None	No
CONNECT	Yes (2013)	2	3	10	10 (2011)	Yes (2011)
COMP	Yes (2013)	2	3	2	None	No
ECFIN	Yes (2008)	2	1	-	-	-
EMPL	Yes (2004)	4.5	3	4	6 (2001)	Yes (2001)
ENER	No	1	1	5	None	No
ENTR	No	2	2	1	12 (2002)	Yes (2002)
ENV	Yes (2007)	1.4	1	20	None	Yes (2003)
ESTAT	No	2	2	-	None	No
HOME	No	1.3	2	5	12 (2011)	Yes (2011)
JUST	Yes (2011)	0.25	1	5	None	Yes (2011)
MARE	No	0.5	2	1	None	No
MARKT	Yes (2008)	1	1	6	15 (2008)	Yes (2008)
MOVE	No	2.5	1	4	None	No
SANCO	No	2	1	3	None	None
TAXUD	Yes (2012)	2.5	1	3	None	No
TRADE	Yes (2011)	1	1	6	3 (2008)	No

As Table 4 shows, nine DGs organized an *evaluation training* in 2014. Most of these trainings were set up during the last few years. Other DGs only participate in the centralized training organized by the SG (five days total), which is compulsory for all evaluation-related staff. Because trainings at the DG-level usually last one day at most, they are more suitable to reach a broad audience of policy makers.

Concerning the *number of fte* working for the evaluation functions, it turned out that in some DGs coordinating evaluations in only a part of the job of a single person (MARE, JUST), while others dedicate a small team to the issue (EMPL, ENTR, TAXUD, MOVE). Staff differences are present in relative terms as well: DG Agriculture has about 1.4% of its staff working on coordinating evaluations, while for DG JUST this is 0.05%.[5] It should be noted, however, that the high numbers for DG AGRI and CONNECT are partly caused by the fact that their evaluation functions conduct some programme evaluations themselves, rather than only supporting other units.

Concerning *evaluation expertise*, all DGs participate in the central evaluation network organized by the SG, but beyond that there is much variation. DG COMP, HOME, MARE and EMPL have internal evaluation networks in which their various directorates are represented, while DG AGRI and EMPL organize networks with member state evaluation experts. DG COMP contributes to an OECD evaluation network, while DG ESTAT, ENTR and CONNECT participate in evaluation-related networks of all DGs working in a specific policy area. Other DGs only work with infrequent (lunch) sessions to discuss evaluation.

Concerning external expertise, most DGs work with framework contracts for their legislative evaluations, meaning that only a limited number of preselected companies (between three and six) may bid for contracts. DG CONNECT and DG ENV usually allow open competition between companies, which explains the high number of average bids they receive. Using open competition can take twice as much time as using framework contracts, which is why most DGs prefer the latter, although most respondents do believe that going to the market offers extra quality.

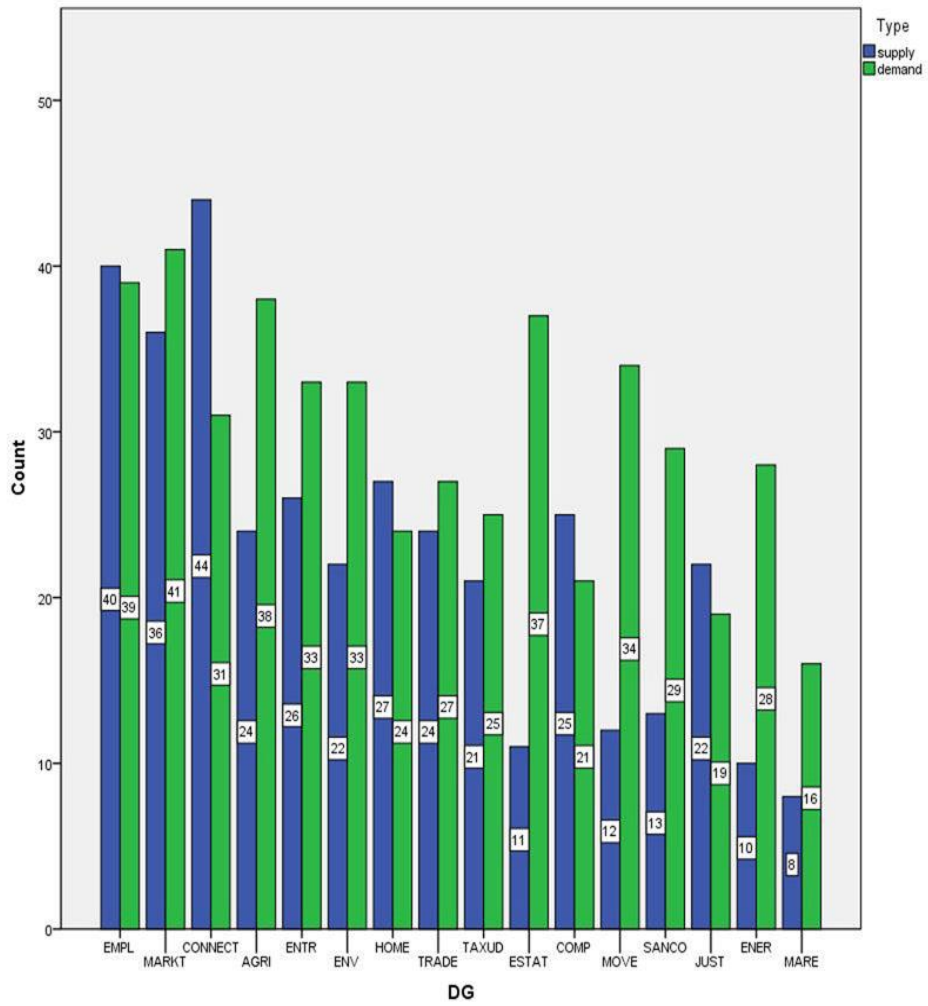
As for *evaluation methods* and *evaluation models*, only DG CONNECT and DG MARKT have guidelines about these topics specifically for legislative evaluation. DG EMPL, ENTR, ENV, HOME, JUST, SANCO and TRADE have internal guidelines for ex-post evaluation in general. While DG MARKT, CONNECT, HOME and ENTR discuss ten or more methods in their guidelines, the guidelines of other DGs discuss only a small number of methods or no methods at all, and the document of DG SANCO discusses neither methods nor models.

A question that remains is whether these results based on 2014 are still up-to-date, as the SG published new evaluation guidelines in May 2015 (European Commission, 2015). None of the respondents believed that the new guidelines would lead to immediate human or financial investments in evaluation at the DG-level. However, as the new guidelines do specify rules for writing follow-up action plans (European Commission, 2015: 297-298), the current variation among DGs on that aspect might be reduced. Furthermore, at the beginning of 2015 DG HOME and DG TRADE have created units dealing specifically with evaluation and related matters to reflect the growing importance of these topics.

Using the relative weights of the indicators listed in Table 2, Figure 1 provides the final scores for each DG on evaluation demand, evaluation supply and total evaluation capacity. DGs

are ranked from highest to lowest, but the scores should be seen as descriptions rather than judgements. The results show a large group of DGs receiving between 45 and 55 points in total, with a few outliers having higher and lower scores. Values for evaluation demand are generally a little higher than for evaluation supply, especially for the DGs at the lower end of the spectrum, indicating that these DGs may deem evaluation important but have little means to invest in it.

Figure 1: scores on evaluation supply and demand for each DG. Values for ESTAT are adapted proportionally to compensate for missing data.



5. Analysis

When applying fsQCA, a useful first step is to test if any individual explanatory conditions are either necessary or sufficient to let the outcome occur. Therefore, Table 5 shows which conditions provide consistent explanations for high evaluation demand and supply (benchmark: >0.80) and which consistency scores above the threshold are probabilistically significant (benchmark: <0.05). Unlike the consistency threshold, the probabilistic test also takes the number of cases in which a condition leads to an outcome into account (Ragin, 2008: 120), which makes it useful for this article because some of the conditions are only present in a handful of cases. Since in fsQCA the absence of an explanatory condition does not necessarily have the opposite consequences of the presence of that condition, the negation of each condition is also included in the table, recognizable by the symbol ~.

As the results show, neither the amount of legislation nor the political sensitivity of a DG nor their negations are necessary or sufficient conditions for high evaluation demand or supply (hypotheses 1, 2, 5 and 6 are falsified). This outcome is also apparent if we look at the four DGs with the highest overall capacity: DG EMPL, CONNECT, MARKT and AGRI. In the first two of these DGs the condition of having a large amount of legislation is clearly absent (both DGs initiated about 25 regulations and directives, while the cut-off point is 50), and out of the four cases only DG EMPL deals with a policy field that is considered politically sensitive.

However, in line with hypothesis 3, the presence of a high budget is a sufficient condition for high evaluation demand. In other words, when a DG has many financial resources we can expect it to attach much importance to legislative evaluation. The four DGs with budgets of more than €1 billion (EMPL, CONNECT, ENTR and AGRI) all have high evaluation demand as well. The corresponding coverage score is 0.44, showing that the high budget condition accounts for a little less than half of the cases with high evaluation demand.

The presence of a high budget is also a sufficient condition for high evaluation supply, in line with hypothesis 4. This relationship does not pass the probabilistic test, but when taking a closer look at the data that fact is explained entirely by the case of DG AGRI, which has a very high budget and only a medium score (24 points) on evaluation supply. Generally speaking, it therefore seems that DGs with a high budget also invest a large amount of human and technological capital in legislative evaluations. The corresponding coverage score is 0.51,

showing that the high budget condition accounts for about half of the cases with high evaluation supply.

According to the theoretical framework of this article, these results indicate that DGs with a tradition of evaluating spending programmes also have high capacity for legislative evaluations. This interpretation is supported by statements made by several respondents during the interviews. For example, one evaluation coordinator from a low-budget DG stated that evaluation used to be a low priority in his organization because the evaluation culture in the Commission was so focused on accounting for how money was spent. Only in the last four years a shift began towards legislative evaluations, and since then his DG has slowly started building evaluation capacity. Another respondents emphasized that his DG already conducts evaluations since the 1990s, starting with spending programmes, and has therefore become a frontrunner concerning all kinds of ex-post evaluations in the Commission.

Table 5: results of QCA analysis for evaluation demand / evaluation supply. The level of significance for all proportions > 0.80 is provided in parenthesis. ~ represents the negation of a condition.

Condition	**Consistency (necessary conditions) for evaluation demand / evaluation supply**	**Consistency (sufficient conditions) for evaluation demand / evaluation supply**
High legislative amount	0.63 / 0.63	0.79 / 0.62
~High legislative amount	0.63 / 0.72	0.69 / 0.61
High budget	0.44 / 0.51	0.95 (0.001)* / 0.84 (0.65)
~High budget	0.85 (0.56) / 0.82 (0.85)	0.70 / 0.51
High sensitivity	0.23 / 0.30	0.54 / 0.55
~High sensitivity	0.78 / 0.70	0.61 / 0.43

Besides looking at individual conditions, fsQCA also allows for analysing combinations of causal conditions (Ragin, 2008: 125). The truth table (Table 6) shows all combinations that appear in the data with their corresponding cases. Only those combinations with a consistency score of more than 0.8 were included in the analysis, to ensure that the results remain undistorted by combinations with contradictory scores on evaluation demand and supply (Ragin, 2008: 139).

To analyse the truth table, the intermediary method was used (Ragin, 2008: 164)[6]. The results show that the absence of political sensitivity in combination with the presence of a large amount of legislation consistently leads to high evaluation demand (consistency = 0.77) and that this solution covers about one-third of the cases with high demand (unique coverage = 0.29). This result indicates that DGs will usually value legislative evaluation if they have a strong legislative responsibility in a policy field which is not so sensitive that evaluations might be threatening. However, more research is needed to confirm if this interpretation is correct.

When including the presence of high budgets, the entire solution for high evaluation demand has a consistency of 0.80 and a coverage of 0.73, meaning that it covers about three-quarters of the cases with high evaluation demand. No combinations of conditions were found which explain the presence of high evaluation supply.

To check the robustness of these findings, various higher and lower cut-off points for the explanatory conditions were tested[7]. The results were largely unaffected by these changes: the presence of a high budget kept being a sufficient condition for high demand (consistency > 0.9), while most other individual conditions remained neither necessary nor sufficient. However, if the cut-off point for high budgets is put below €775 million, which is close to the medium-sized budget of a case like DG MARE, it seizes to be a sufficient condition for high evaluation supply.

The analysis so far focused on explaining the presence of high evaluation demand and supply, which in fsQCA is not the same as explaining the absence of these outcomes (Ragin, 2008: 102). A similar analysis was therefore conducted for the negations of the outcomes. Its results cannot be fully presented here due to word constraints, but it can be said that its results were in line with the previous findings. The absence of high budgets turned out to be a necessary condition for both low evaluation demand (consistency = 0.96; α = 0.00) and low evaluation supply (consistency = 0.92; α = 0.05), indicating that almost all DGs with low evaluation capacity also have low budgets (DG AGRI being the only exception). No other individual conditions consistently explained the absence of the outcomes, nor did any combinations of conditions.

Table 6: truth table

High amount of legislation	High budget	High sensitivity	N	Cases	Outcome (high demand)	Outcome (high supply)	Raw consistency (demand / supply)
No	No	No	5	COMP, ENER, ESTAT, MARE, TRADE	No	No	0.69 / 0.48
Yes	No	No	4	ENV, MARKT, MOVE, SANCO	Yes	No	0.83 / 0.60
Yes	Yes	No	2	AGRI, ENTR	Yes	Yes	1.00 / 0.95
No	No	Yes	2	JUST, TAXUD	No	No	0.72 / 0.72
No	Yes	Yes	1	EMPL	Yes	Yes	0.98 / 1.00
No	Yes	No	1	CONNECT	Yes	Yes	0.99 / 0.88

6. Conclusion

This article addressed the questions how the DGs of the European Commission vary in their capacity for legislative evaluations and how this variance can be explained. Through in-depth interviews with twenty Commission officials, data were collected about both evaluation supply and demand. The results reveal much variance in capacity between DGs. On the demand side, some DGs have very clear aims and procedures for all the stages of legislative evaluation, while for other DGs this is not the case. On the supply side, while for some DGs coordinating evaluation is a part-time job of one person, others devote a small team or even a whole unit to the task. Over the last few years the number of DGs supporting their staff with their own evaluation-related trainings, networks and guidelines has gradually increased, but each of these features is still present in only about half of the DGs. The highest overall evaluation capacity for 2014 was found in DG EMPL, MARKT, CONNECT and AGRI.

How can this variance be explained? The analysis shows that the presence of a high budget is a sufficient condition for high evaluation demand and supply, indicating that DGs with a long tradition of evaluating spending programmes attach more importance to and invest more means in legislative evaluations than other DGs. Theoretically there could be other explanations for the relationship between high budgets and high capacity, but the qualitative information from the interviews confirms that DGs with a long tradition of evaluating spending activities also

pay more attention to legislative evaluations today. Furthermore, the analysis indicates that DGs with have a strong legislative responsibility in a policy field which is not very sensitive usually have high evaluation demand. However, more qualitative research should be conducted to verify this interpretation.

Two other possibilities for future research stand out. First, this article was mostly focused on the presence of certain minimal organizational requirements to systematize legislative evaluation in the Commission's DGs. It has not studied the extent to which the tools and procedures available for legislative evaluations are applied in practice by the DGs, nor the question for which aims legislative evaluations are used in the Commission. These topics would be worthy of further investigation.

Second, future research could take a look at the consequences of capacity differences for the initiation and the quality of legislative evaluations. From a technical perspective we could expect that DGs which invest more human and technological capital in legislative evaluations (i.e. DGs with high evaluation supply) produce more and better evaluations. From a more political perspective, we could expect that DGs which attach more value to evaluation (i.e. DGs with high evaluation demand) show better evaluation outputs (Mastenbroek et al., 2016). By further studying these topics, the data on evaluation capacity presented in this article could help to enhance our understanding of both the nature of the European Commission and the role of legislative evaluations in the European policy cycle.

Notes

[1] This concerned the DGs ESTAT (phone), ECFIN (phone) and MOVE (e-mail).

[2] Respondents from DG ENTR, MARKT and MOVE were unable to provide average evaluation costs, so for these DGs this information was collected by taking the average of three cost indications mentioned in tender or contract specifications published at http://ec.europa.eu/transport/facts-fundings/tenders/index_en.htm

[3] Annual activity reports for 2013-2014 are available at http://ec.europa.eu/atwork/synthesis/aar-archived/aar_2013_en.htm. Activities categorized as both spending and non-spending activities were counted as half a spending activity.

[4] The analysis was mostly conducted with the fuzzy add-on in Stata. For details about this add-on, see http://www.stata-journal.com/sjpdf.html?articlenum=st0140

[5] Based on 2015 staff figures: http://ec.europa.eu/civil_service/docs/europa_sp2_bs_cat-sexe_x_dg_en.pdf

[6] The intermediary solution, which is the recommended approach for fsQCA is most circumstances (Ragin, 2008: 164), only takes configurations which do not appear in the data into account if they meet certain assumptions. Following the theoretical framework of the article, positive relations were assumed between high amounts of

legislation and high budgets and the outcomes, while negative relations were assumed between high political sensitivity and the outcomes.

[7] The cut-off point for the amount of legislation was decreased and increased by up to twenty pieces of legislation; the cut-off point for budgets was decreased and increased by up to €500 million. In both cases, the most extreme cut-off points tested left only a few cases above or below their values.

References

Borrás S and Højlund S (2015) Evaluation and policy learning: The learners' perspective. *European Journal of Political Research* 54(1): 99-120.

Bourgeois I and Cousins JB (2013) Understanding Dimensions of Organizational Evaluation Capacity. *American Journal of Evaluation* 34(3): 299-319.

Bovens M, Hart P 't and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford handbook of public policy*. Oxford: university press, pp. 320-335.

Boyle R, Lemaire D and Rist RC (1999) Introduction: Building evaluation capacity. In: Boyle R and Lemaire D (eds) *Building effective evaluation capacity: Lessons from practice*. New Brunswick USA: Transaction Publishers, pp. 1-19.

Carman JG and Fredericks KA (2010) Evaluation Capacity and Nonprofit Organizations. Is the Glass Half-Empty or Half-Full? *American Journal of Evaluation* 31(1): 84-104.

European Commission (2007) *Communication from the Commission from ms Grybauskaité in agreement with the president. Responding to Strategic Needs: Reinforcing the use of evaluation [SEC(2007)213].* Brussels: European Commission.

European Commission (2013) *Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions. Strengthening the foundations of smart regulation: improving evaluation [COM(2013)686]*. Brussels: European Commission.

European Commission (2015) *Better Regulation Toolbox [complement to SWD(2015)111]*. Brussels: European Commission.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

Højlund S (2015) Evaluation in the European Commission - For accountability or learning? *European Journal of Risk Regulation* 6(1): 35-46.

Kogut B, McDuffie JP and Ragin CC (2004) Prototypes and strategy: assigning causal credit using fuzzy sets. *European Management Review* 1(2): 114-131.

Majone G (1999) The regulatory state and its legitimacy problems. *West European Politics* 22(1): 1-24.

Mastenbroek E, Van Voorst S and Meuwese ACM (2016) Closing the regulatory cycle? A meta-evaluation of ex-post legislative evaluations by the European Commission. Epub ahead of print 5 October 2015. DOI: 10.1080/13501763.2015.1076874

Nielsen SB, Lemire S and Skov M (2011) Measuring evaluation capacity: Results and implications of a Danish study. *American Journal of Evaluation* 32(3): 324-344.

Nugent N (2010) *The government and politics of the European Union (7th edition)*. Houndmills UK: Palgrave.

Pattyn V (2014) Why organizations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation* 20(3): 348-367.

Preskill H and Boyle S (2008) A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation* 29(4): 443-459.

Ragin CC (2008) *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: university press.

Ramboll management consulting (2011) *The evaluation capacity index*. Available on request from EvaluationSociety@r-m.com.

Scharpf FW (1999) *Governing in Europe: Effective and democratic?* Oxford: University Press.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation practice: New directions for evaluation*. San Fransisco CA: Jossey-Bass, pp. 67-85.

Stockdill SH, Baizerman M and Compton DW (2002) Toward a definition of the ECB process: A conversation with the ECB literature. *New Directions for Evaluation* 93: 7-25.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation.* New Brunswick: Transaction, pp. 407-424.

Taut S (2007) Studying self-evaluation capacity building in a large international development organization. *American Journal of Evaluation* 28(1): 45-59.

Taylor-Ritzler T, Suarez-Balcazar Y, Garcia-Iriarte E, Henry DB and Balcazer FE (2013)

Understanding and Measuring Evaluation Capacity: A Model and Instrument Validation Study. *American Journal of Evaluation* 34(2): 190-206.

Technopolis (2005) *Study on the use of evaluation results in the European commission*. Brussels: European Commission.

Toulemonde J, Summa-Polit H and Usher N (2005) Triple check for top quality or triple burden? Assessing EU evaluations. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction, pp. 66-90.

Weiss CH (1993) Where Politics and Evaluation Research Meet. *American Journal of Evaluation* 14(1): 93-106.

Chapter 4: Enforcement tool or strategic instrument? The initiation of ex-post legislative evaluations by the European Commission

Stijn van Voorst and Ellen Mastenbroek

Published as: Van Voorst S and Mastenbroek E (2017) Enforcement tool or strategic instrument? The initiation of ex-post legislative evaluations by the European Commission. *European Union Politics* 17(4): 640-657.

Abstract

Whereas the European Commission officially intends to periodically evaluate all major EU legislation in force, in practice it only evaluates a minority of major regulations and directives. This article tries to explain the variation in the initiation of such ex-post legislative evaluations by the Commission with the help of two theoretical motives: an enforcement motive and a strategic motive. Based on two novel datasets and binary logistic regression analysis, the results show that the type and complexity of the legislation, the presence of an evaluation clause and the evaluation capacity of the responsible DG enhance the chances of evaluation. These findings indicate that ex-post legislative evaluations are at least partly driven by the Commission's need to enforce legislation.

1. Introduction

The European Union (EU) is often described as a 'regulatory state' due to the important role of legislation in the European policy process (Majone, 1999: 1). A marked feature of the European legislative process is the centrality of one supranational executive actor: the European Commission. The Commission has a number of crucial tasks related to European legislation. Firstly, it is responsible for the development and formulation of legislative proposals (Schmidt

and Wonka, 2013: 2). Secondly, it produces delegated and implementing acts (McCormick, 2015: 169-72). Thirdly, in its role of 'guardian of the European treaties', the Commission is responsible for monitoring and enforcing national compliance with European legislation (McCormick, 2015: 169-72; Schmidt and Wonka, 2013: 2).

These three tasks of the Commission in the EU's legislative process have received ample academic scrutiny (e.g. Kassim et al., 2013; Schmidt and Wonka, 2013; Wille, 2013). Conversely, the literature has hardly touched upon a fourth key task of the Commission, which is to conduct ex-post evaluations that assess the functioning and effectiveness of European legislation. So far, such ex-post legislative (EPL) evaluations have mostly been neglected by scholars (but see Fitzpatrick, 2012; Mastenbroek et al., 2016; Zwaan et al., 2016), which is all the more surprising given both their theoretical importance and their growing role in the Commission.

Theoretically speaking, EPL evaluations may fulfil two important functions in legislative processes. Firstly, by recommending how the implementation of legislation can be improved and/or how legislation can be amended to increase its effectiveness, EPL evaluations are a potential tool for decision-makers to improve their policies (Fitzpatrick, 2012: 479; Vedung, 1997: 109). Secondly, EPL evaluations can be used to judge the performance of the actors that implement legislation, thus holding them accountable for their actions (Coglianese, 2012: 11; Vedung, 1997: 102-108).

Over the years the Commission has increasingly recognized the importance of EPL evaluations. It first emphasized the role of such evaluations in legislative improvement and accountability relationships in 2000, after which it started to make its procedures for EPL evaluations more systematic in 2007 (European Commission, 2007: 3-4; Fitzpatrick, 2012: 478). Since 2010 the Commission has also stressed the importance of EPL evaluations for judging the suitability of entire regulatory frameworks (so-called 'fitness checks') (European Commission, 2010: 5). Furthermore, from 2012 onwards it has given EPL evaluations a central place in its REFIT programme, which aims to identify and remove superfluous rules (European Commission, 2012: 4).

In 2015 the Commission published new guidelines that outline the methods, follow-up procedures and institutional responsibilities for carrying out EPL evaluations (European Commission, 2015). In principle, all the Commission's EPL evaluations must use some form of

stakeholder consultation to map the views of those actors that are directly affected by EU legislation (European Commission, 2015: 299-336). Aside from this, EPL evaluations can use different combinations of methods, such as expert interviews, document analysis and quantitative modelling (European Commission, 2015: 337-414; Fitzpatrick, 2012: 490-7).

Concerning the follow-up of EPL evaluations, the Commission is supposed to produce an action plan based on the main recommendations of each evaluation to ensure that its results feed back into the 'regulatory cycle' (European Commission, 2015: 297-298). Existing research has shown that the extent to which this happens varies in practice, such that about half of the ex-ante evaluations (impact assessments) attached to proposals for legislative amendments make use of information from EPL evaluations when available (Van Golen and Van Voorst, 2015: 388).

Concerning the institutional responsibility for EPL evaluations, the Commission's guidelines specify that such evaluations are the responsibility of the Directorates-General (DGs), with a coordinating role for the Commission's Secretariat-General (SG) (European Commission, 2015: 257; Stern, 2009: 70-71). EPL evaluations are usually based on reports written by external consultants to enhance their independence, but when this is more practical the whole evaluation process may also be conducted internally (European Commission, 2015: 282-9).

Importantly, since 2007 the Commission's guidelines also prescribe that both financial and legislative activities must be evaluated periodically, in proportion to their allocated resources and expected impact (European Commission, 2007: 22; 2015: 257). In reality, however, not all important EU legislation is evaluated. Academic research has shown an initiation ratio of 33% for major EU regulations and directives from the period 2000-2012 (Mastenbroek et al., 2016: 1338). The Commission (2013: 13) has produced similar figures: in 2013, 29% of all important EU regulations had been evaluated, with a further 13% of these regulations being evaluated at that moment, 19% of these regulations having a future evaluation planned and no numbers being provided for directives.

These figures show that the Commission is apparently selective in which legislation it evaluates, for reasons that the institution itself does not explain. This finding is problematic because an evaluation system is only credible if its procedures for initiating evaluations are systematic and transparent (OECD, 2015: 120). If this is not the case, legislative quality may

diminish in policy areas that are evaluated less frequently (OECD, 2015: 120) and/or the image could arise that the Commission decides what legislation to evaluate based on political considerations (Radaelli and Meuwese, 2010: 146). This, in turn, could harm the credibility of evaluations in the eyes of the legislator and other actors (Poptcheva, 2013: 4).

Therefore, this article looks into the question of what drives the initiation of EPL evaluations by the Commission. In other words, why does the Commission evaluate some pieces of law while it does not evaluate others? By answering this question, we not only seek to shed light on the unexplored topic of EPL evaluations in the EU, but also aim to further explore the motives that drive the Commission's behaviour (Boswell, 2008: 472; Franchino, 2007: 11; Hartlapp et al., 2014: 1; Radaelli, 1999: 760-2; Wille, 2010: 1098-1100).

Two potential motives (not) to initiate an evaluation are studied in this article, each of which is linked to a specific theoretical image of the Commission. The first motive, which is in line with the image of the Commission as the 'guardian of the European treaties', is the effective enforcement of EU legislation. Since EPL evaluations are a potential tool to check how legislation is implemented by the member states (European Commission, 2015: 296; Stame, 2008: 124), we can expect that legislation for which the chances of non-compliance are higher is more likely to be evaluated. The second motive is the strategic protection of competences, which is in line with the image of the Commission as a political actor (Boswell, 2008: 472; Hartlapp et al., 2014: 1; Majone, 2005: 65). Following this logic, we would expect that the Commission refrains from evaluating legislation if this could result in a reduction of its powers.

The hypotheses flowing from these two motives are tested with the help of two datasets, the first containing all major EU legislation from 2000-2004 and the second containing all EPL evaluations conducted by the Commission during 2000-2014. With these data, we are able to draw conclusions about the decisions to evaluate European legislation over a 15-year period. The 10-year gap between our datasets is needed to give the Commission enough time to evaluate, thus avoiding any bias in our data in favour of legislation that was evaluated sooner. Binary logistic regression was used for the analysis.

Our results show that EU legislation is more likely to be evaluated if it is a directive rather than a regulation and if it is more complex, which is in line with the enforcement motive. Both of our control variables - the presence of evaluation clauses and the amount of evaluation

capacity of the DG to which a piece of law belongs - also provide significant explanations. However, we did not find evidence that the strategic protection of competences explains the Commission's initiation of EPL evaluations.

2. Theoretical framework

Whereas evaluation-related topics are frequently discussed in the academic literature, there is no comprehensive approach to explaining why organizations decide to evaluate or not (Mastenbroek et al., 2016: 1343; Pattyn, 2014: 351). Therefore, this article develops such an approach in the context of the EU, building on two potential motives for the Commission: an enforcement motive and a strategic motive. These motives are closely linked to ongoing academic debates about the nature of the Commission (Boswell, 2008: 472; Franchino, 2007: 11; Hartlapp et al., 2014: 1; Radaelli, 1999: 760-2; Wille, 2010: 1098).

Enforcement motive

In its role of 'guardian of the European treaties', the Commission has the task to monitor and enforce member state compliance with EU legislation (Schmidt and Wonka, 2013: 2). EPL evaluations are potentially useful for this purpose, as they can collect and present information about how rules are implemented in practice (Coglianese, 2012: 11). This, in turn, makes EPL evaluations useful to hold those actors responsible for the implementation of legislation accountable (Vedung, 1997: 102). Therefore, EPL evaluations are a potential tool for the Commission to detect non-compliance by the member states and to address such non-compliance via enforcement measures (European Commission, 2015: 292; Stame, 2008: 124).

The role of EPL evaluations in enforcing European legislation is also evident from earlier research about this topic. Mastenbroek et al. (2016: 1339) found that out of 216 EPL evaluations conducted or outsourced by the Commission between 2000 and 2012, 79% assessed the processes of legislative implementation, enforcement and/or compliance. Zhelyazkova et al (2016: 833) found EPL evaluations to be the most detailed source of information about the compliance of member states with 24 directives of interest.

Those EPL evaluations that study member state compliance often assess the legal implementation of directives by systematically comparing national transposition measures,

while they tend to assess the practical implementation of European legislation via surveys and interviews among stakeholders. In some cases, infringement data are also used as a source (Smith, 2015: 92-93). EPL evaluations that address member state compliance also tend to include recommendations for the Commission. Often these recommendations focus on 'soft' measures like increased monitoring, sharing best practices or publishing guidelines for national implementing authorities, but evaluations may also recommend the Commission to launch infringement procedures (Mastenbroek et al., 2016).

If the Commission can use EPL evaluations for enforcement purposes, we can expect that the chances than an evaluation is initiated are higher for pieces of law where there is a greater need to scrutinize the member states. In other words, we can expect that the chances that an evaluation is initiated are higher for legislation that offers more opportunities for non-compliance.

Three specific variables may be important in this regard. Firstly, the *type of legislation* may affect the chances of non-compliance. Directives offer the member states more discretion than regulations because they need to be transposed into national law (Treib, 2014: 6). In turn, this discretion offers the member states more opportunities to delay or prevent implementation (Kaeding, 2006: 232; König and Mäder, 2014: 247; Mastenbroek, 2003: 372; Steunenberg and Rhinard, 2010: 495; Treib, 2014: 6). Therefore, we expect directives to be more likely to be evaluated than regulations (Stame, 2008: 124).

Hypothesis 1: Directives are more likely to be evaluated than regulations.

Secondly, the *complexity of legislation* can affect the chances of non-compliance. Since the European legislative process includes multiple veto players - notably the Commission, the Council and the EP - decision making often produces compromises that are laid down in long and ambiguous texts (Häge, 2007: 307-308; Hofmann, 2013: 99). Such complexity offers member states more leeway for interpretation, and, therefore, makes it more difficult to establish whether they are complying with legislation or not (Kaeding, 2006: 242; König and Mäder, 2014: 253-254; Mastenbroek, 2003: 376; Steunenberg and Rhinard, 2010: 501). This in turn can be expected to increase the chance that legislation is evaluated by the Commission.

Hypothesis 2: The more complex a piece of law, the higher its chances of being evaluated.

Thirdly, the political sensitivity of legislation may affect the chances of non-compliance. The more controversial a regulation or directive, the more likely it is that some member states who opposed it during the legislative process will not implement it correctly (Mastenbroek, 2003: 376). Since the Council represents the member states, *politicization in the Council* is especially likely to increase the chances of non-compliance (Treib, 2014: 14), and, therefore, the chances that an evaluation is initiated.

Hypothesis 3: The more politicized a piece of law was in the Council; the more likely it is to be evaluated.

Strategic motive

The hypotheses presented above are in line with the image of the Commission as a 'guardian of the European treaties'. However, in recent years scholars have increasingly viewed the Commission as an actor that not only fulfils the tasks that the member states have delegated to it (such as enforcing European legislation), but also strategically pursues its own preferences (Franchino, 2007: 11; Hartlapp, 2014: 1-14; Wille, 2010: 1099). According to this political view on the Commission the institution has a perpetual interest to protect its competences, as without these competences it would not be able to achieve any of its (temporary) political aims (Hartlapp, 2014: 1; Majone, 2005: 65; Pollack, 2008: 9).

The Commission has been shown to deal strategically with ex-ante evaluations of legislation (impact assessments) (Poptcheva, 2013: 4; Torriti, 2010: 1065) and expert knowledge in general (Boswell, 2008: 472), so we can expect strategic considerations to play a role in decisions about the initiation of EPL evaluations as well. Ex-post evaluations are not just neutral instruments that can be used to stay informed about policy implementation, but also potential strategic tools that can strengthen or weaken the positions of actors (Bovens et al., 2008: 320; Schwartz, 1998: 295; Vedung, 1997: 111). As evaluations suggest changes to existing arrangements, they are inherently advantageous to some actors and disadvantageous to others (Bovens et al., 2008: 320; Weiss, 1993: 95-98). Negative evaluations can be particularly

disadvantageous to actors that are responsible for delivering policies, as such evaluations may lead to demands to roll back their competences or to put them under closer supervision (Vedung, 1997: 102-8). This, in turn, may be an incentive for such actors to avoid evaluations that may have negative consequences (Schwartz, 1998: 295, Weiss, 1993: 95).

Therefore, we can expect the Commission to be reluctant to initiate EPL evaluations in situations where the results of such evaluations could be harmful to its interests. The Commission's better regulation agenda officially endorses the idea that EU legislation should be significantly amended or even repealed if an evaluation shows that it has no added value (European Commission, 2012: 3; 2013: 1; 2015: 254). In reality, however, we can expect that the Commission wants to avoid such situations to protect its competences (Majone, 2005: 65). In other words, we expect the chances that a piece of law is evaluated to be lower if the potential evaluation is more likely to be used to argue for significant amendments to the law.

Involvement of the European Parliament (EP) in decision making decreases the chances of significant amendments and is therefore expected to increase the chances that an evaluation is initiated. The reason for this is that the EP provides an extra veto player that can block amendments (Häge, 2007: 307; Hofmann, 2013: 102). As a majority of EP members generally supports further European integration (Pollack, 2008: 9), it can also be expected that the EP will usually oppose reducing the competences of supranational institutions like the Commission.

Hypothesis 4: Pieces of law that can only be amended with the approval of the European Parliament are more likely to be evaluated than pieces of law that can be amended without the approval of the European Parliament.

The *voting procedure in the Council* is also expected to influence the chances of legislative amendments. If unanimity is required in the Council it is significantly harder to change legislation, as it is difficult to make all member states agree on a proposal (Häge, 2007: 308). We therefore expect the chances that an evaluation is initiated to be higher when the Council applies unanimity voting, as compared to when it applies qualified majority voting (QMV).

Hypothesis 5: Pieces of law decided upon by unanimity in the Council are more likely to be evaluated than pieces of law decided upon by qualified majority voting.

Control variables

Aside from the two theoretical explanations described above, this research controls for two other potential explanations for decisions to initiate EPL evaluations. The first control variable is the *presence of an evaluation clause*. Many EU regulations and directives contain provisions that oblige the Commission to evaluate them after a number of years, which are usually inserted by the Council and the EP to ensure that they will stay informed about the legislation (Summa and Toulemonde, 2002: 410). We can expect legislation containing an evaluation clause to be evaluated more often than legislation without such a clause.

The second control variable is the *evaluation capacity of the responsible DG*. In this context, evaluation capacity is defined as the presence of sufficient means and processes to ensure that evaluation is an ongoing practice in an organization (Nielsen et al., 2011: 325). Evaluation capacity includes the presence of organizational structures and procedures that support evaluations, the presence of sufficient financial and human capital to evaluate and the presence of proper (methodological) tools to conduct evaluations (Nielsen et al., 2011: 326-7). Since evaluation capacity varies primarily between the DGs of the Commission (Van Voorst, 2017: 25), we expect that legislation under the responsibility of DGs with higher evaluation capacity is more likely to be evaluated than other legislation.

We could also expect the Commission to be more likely to initiate an EPL evaluation if an ex-ante evaluation (impact assessment) of the same legislation was carried out, as impact assessments often contain a section prescribing that legislation should be evaluated ex-post (European Commission, 2015: 246-51). However, this variable cannot be studied in this research because the Commission's system for impact assessments was only set up in 2002-2003 (European Commission, 2007: 4).

3. Methods and data

Data collection

Although multiple datasets of EU legislation already exist (e.g. Hofmann, 2013: 102; Treib, 2014: 27), none of them suited the specific aims of our research. Therefore, we created two datasets for the task at hand, one containing major European legislation and one containing EPL evaluations (also see chapter 2 of this dissertation).

The dataset of legislation covers the years 2000-2004. This period was chosen to give the Commission sufficient time to evaluate. While academic literature indicates that legislation is usually evaluated after about five years (Eijlander and Voermans, 2000: 355) and evaluation clauses in EU legislation also tend to give the Commission five years or less to evaluate, we decided to double this period to avoid concluding that any legislation has not been evaluated while an evaluation was in fact still upcoming. Therefore, our dataset of legislation stops at the end of 2004, but it should be emphasized that our article concerns decisions to initiate EPL evaluations over a period of fifteen years (2000-2014), which is further explained by the description of our second dataset below.

Because the Commission follows the logic that the resources spent on an evaluation must be proportionate to the importance of a measure (European Commission, 2007: 22; 2015: 255-6), minor EU legislation does not have to be evaluated (European Commission, 2015: 253). Therefore, our dataset of legislation only includes major regulations and directives. We excluded all delegated and implementing acts,[1] which are generally considered less important than primary legislation (Franchino, 2007: 80), as well as all rectifications, amendments and secondary Council legislation. Because of the explicit link between evaluations and improving the effects of legislation on European citizens and companies (European Commission, 2007: 3; 2012: 2; 2013: 1-2), we also excluded legislation without direct relevance for national actors. This includes legislation that only addresses EU institutions or foreign countries. Together, the selection criteria led to a dataset of 277 major directives and regulations adopted in the period 2000-2004. Our dataset of evaluations (see below) contains only eight evaluations of legislation that we did not consider 'major' (2% of all evaluations), indicating that our selection criteria were fairly appropriate.

To assess the initiation ratio of evaluations, our dependent variable, we extracted information from a second dataset. This dataset contains 313 EPL evaluations of regulations, directives and treaty articles conducted or outsourced by the Commission between 2000 and 2014 (updated version of the dataset described in chapter 2 of this dissertation).[2] Evaluations completed before 2000 were omitted because of a lack of data, and evaluations merely studying prescriptions for foreign countries and EU institutions were excluded for the same reasons as discussed above. We also discarded those evaluation reports that merely summarize other evaluations.

The evaluations were gathered from different sources: The Commission's multi-annual evaluation overview (2010), the Commission's search engine for evaluations,[3] the Commission's work programmes,[4] EU bookshop,[5] annexes to Commission's financial reports,[6] and lists of evaluations found on the websites of DGs. We checked our data using an existing list of evaluations produced by expertise centre Eureval, by running Google searches for evaluations of all major legislation adopted between 1996 and 2010, by searching for background documents of legislation in Eur-lex,[7] and by discussing our data-gathering method with the SG (for a further description of the dataset of EPL evaluations, see chapter 2 of this dissertation).

Operationalization

Starting with the enforcement motive, the *type of legislation* (hypothesis 1) was measured as a dichotomous variable (directive or regulation). The *complexity of legislation* (hypothesis 2) was measured through its number of recitals, as more complex legislation generally requires a larger number of explanations (Franchino, 2000: 74; Kaeding, 2006: 236; Steunenberg and Rhinard, 2010: 501; Treib, 2014: 26). *Politicization in the Council* (hypothesis 3) was measured by determining if a legislative proposal was on the Council's agenda as a *B-point*, as B-points represent the topics that are actively debated at the political level (Häge, 2007: 303; Hofmann, 2013: 126; König, 2008: 149).

Concerning the strategic motive, *involvement of the European Parliament* (hypothesis 4) was measured by looking at the formal procedure used to enact the legislation as stated by Eur-lex. In case of the ordinary legislative procedure (former codecision and cooperation procedures) this involvement was considered high, while in case of the consultation procedure

it was considered low (Häge, 2007: 316). The *voting procedure in the Council* (hypothesis 5) was also measured as a dichotomous variable (QMV or unanimity) using Eur-lex.

Concerning the control variables, we searched each piece of law using specific keywords[8] to establish the *presence of an evaluation clause* (yes/no). We also checked the last five articles of each regulation or directive, as this is the most common place for evaluation clauses. Concerning evaluation capacity, we measured twelve indicators derived from a model developed by Nielsen et al. (2011: 326-30) via interviews with the European Commission (Van Voorst, 2017: 29-31). However, only the presence of a specialized evaluation (sub-)unit (yes/no) and the presence of evaluation guidelines (yes/no) could be established per DG per year and were, therefore, useful as indicators. For legislation that has been evaluated, the data used concern the year when the evaluation was published. For legislation that has not been evaluated, we assumed that this decision was made five years after the legislation was published (the modal value of the time between publication dates of legislation and publication dates of evaluations is six years in our dataset, from which we subtracted one year as evaluations usually take that long to conduct) and determined the scores for evaluation capacity accordingly. However, because this assumption of five years is somewhat arbitrary we also experimented with other time periods, which affected our results to some extent.[9]

The operationalization of all our variables is summarized in Table 1. Because of the binary nature of our dependent variable, logistic regression was used for the analysis. The variables belonging to the two motives to evaluate were entered as blocs to allow for comparisons between the models.

Table 1: Operationalization

Type of variable	Variable and hypothesis number	Indicator	Descriptive statistics
Dependent variable	Evaluation initiated	0 = no evaluation 1 = at least one evaluation	0 = 161 cases 1 = 116 cases
Enforcement motive	Legislation type (H1)	0 = regulation 1 = directive	0 = 136 cases 1 = 141 cases
	Complexity (H2)	Number of recitals	Mean = 22.2 σ = 13.1 Range = 4-73
	Politicization Council (H3)	0 = not discussed as B-point 1 = discussed as B-point	0 = 100 cases 1 = 177 cases
Strategic motive	EP involvement (H4)	0 = consultation procedure 1 = ordinary legislative procedure	0 = 111 cases 1 = 166 cases
	Council voting procedure (H5)	0 = QMV 1 = unanimity	0 = 242 cases 1 = 35 cases
Control variables	Evaluation clause	0 = no evaluation clause 1 = evaluation clause present	0 = 112 cases 1 = 165 cases
	Evaluation capacity	0 = no evaluation unit 1 = evaluation unit present	0 = 186 cases 1 = 96 cases
		0 = no evaluation guidelines 1 = evaluation guidelines present	0 = 166 cases 1 = 111 cases

4. Results

Out of the 277 major regulations and directives in our dataset, 116 have been evaluated ex-post. This is an initiation ratio of 41.9%, meaning that about six out of ten major pieces of EU law from 2000-2004 have not (yet) been evaluated by the Commission. This initiation ratio is higher than the 33% found during earlier research about major legislation from 2000-2002

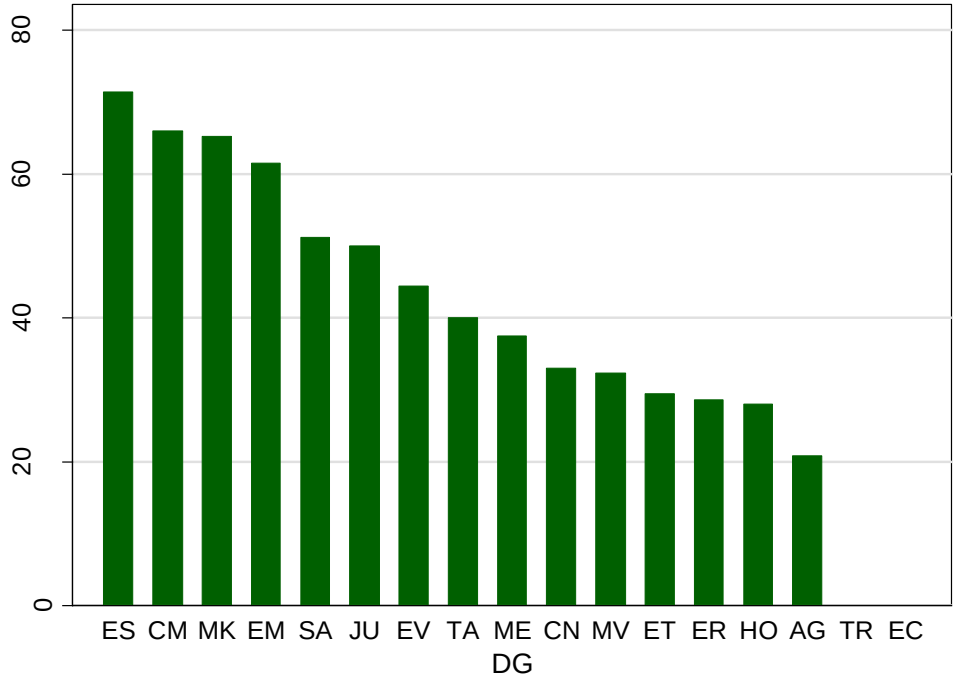
(Mastenbroek et al., 2016: 1338), indicating that legislation published during 2003-2004 was evaluated more often than older legislation. This is a sign that the proportion of legislation evaluated by the Commission may be increasing over time.

A few pieces of law in our dataset were evaluated multiple times over the years: fifteen pieces of law were evaluated twice, four pieces of law were evaluated thrice and two pieces of law were evaluated four times. Due to the binary nature of our dependent variable, these pieces of law with more than one evaluation have no special impact on our analysis: they were simply coded as 1. Their number was also too low to conduct an additional analysis of the number of times that a piece of law was evaluated.

Figure 1 depicts the initiation ratio per DG. The three DGs with the highest initiation ratios are DG Eurostat (71.4%), DG Competition (66.0%) and DG Internal Market (65.2%). DG Trade and DG Economic and Financial Affairs have not evaluated their few major pieces of law from 2000-2004 at all; besides this the three DGs with the lowest initiation ratios are DG Energy (28.6%), DG Home Affairs (28.0%), and DG Agriculture (20.8%). The variation among DGs is included in the analysis through the evaluation capacity variables; the data do not suggest other patterns concerning the size or policy areas of the DGs that warrant investigation.

Table 2 presents the explanatory analysis. The results show that the model with the variables belonging to the enforcement motive and the control variables only passes the chi-square test and is therefore significant. Aside from politicization in the Council all the individual variables included in this model are also significant. Table 2 furthermore shows that if the variables related to the strategic motive are added, the model as a whole and the individual variables that were significant before are still significant. However, neither the involvement of the EP nor the voting procedure in the Council provides an explanation for variation in the initiation of EPL evaluations. Below each variable will be discussed in detail.

Figure 1: Initiation ratio per DG



Notes: Legend of DGs: AG = Agriculture, CM = Competition, CN = Communications and Technology, EC = Economic and Financial Affairs, EM = Employment, ER = Energy, ES = Eurostat, ET = Enterprise and Industry, EV = Environment, HO = Home Affairs, JU = Justice, ME = Maritime Affairs, MK = Internal Market, MV = Transport, SA = Health and Consumers, TA = Taxation, TR = Trade. Some DGs have merged and/or changed their names since 2014.

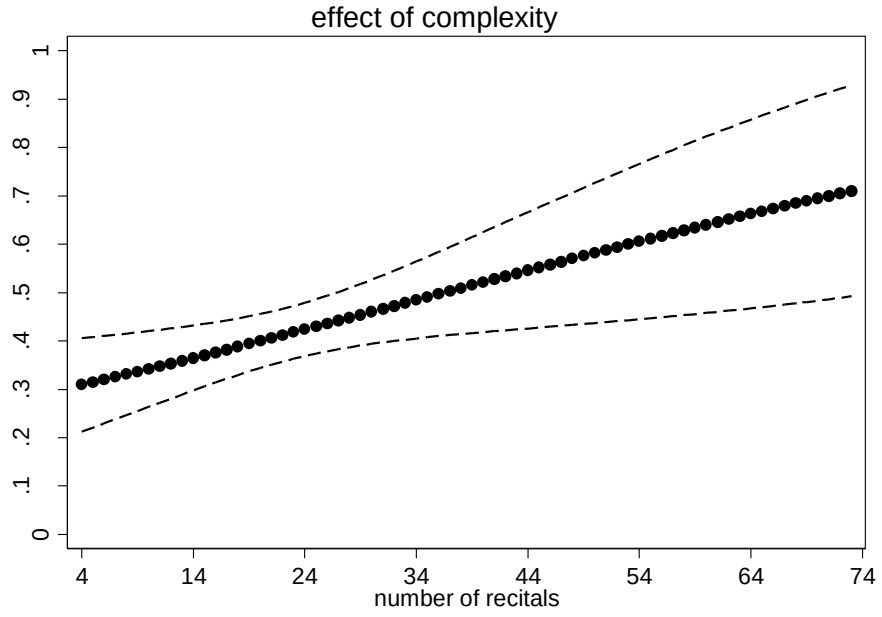
Starting with the *type of legislation*, the first variable belonging to the enforcement model, Table 2 shows that the odds of being evaluated are about 2.05 times higher for directives than for regulations. In terms of predicted probabilities - which are easier to interpret than odds ratios - the chances of an evaluation taking place are 14.0% higher for directives than for regulations, if all other variables are kept at their observed values. In terms of descriptive statistics, out of the 141 major directives in our dataset, 51.8% have been evaluated; out of the 136 major regulations, only 31.6% have been evaluated. In line with hypothesis 1, these findings indicate that the Commission prioritizes evaluating directives over regulations.

The *complexity of legislation*, the second variable belonging to the enforcement model, also significantly increases the chances of an evaluation occurring. For every extra recital, the odds of a piece of law being evaluated increase by about 3%. Figure 2 presents the effect of this variable in terms of predicted probabilities. The figure shows that the chances of an evaluation occurring increase from about 0.3 to 0.7 as the number of recitals grows, with an average growth in predicted probability of 0.06% per recital, if all other variables are kept at their observed values. These findings are in line with hypothesis 2.

Politicization in the Council, the third variable belonging to the enforcement model that was measured by the occurrence of the legislative proposal as a B-point on the Council's agenda, is not significant. Accordingly, we reject hypothesis 3.

Turning to the political variables, the results in Table 2 show that neither the involvement of the EP nor the voting procedures in the Council provides a significant explanation for variation in the initiation of EPL evaluations. This means that hypotheses 4 and 5 are rejected. These results indicate that the chances of a piece of law being significantly amended do not affect the Commission's decision to evaluate this legislation or not.

Figure 2: Effect of legislative complexity on the probability of an evaluation occuring



Conversely, both control variables turn out to be significant. Table 2 shows that the odds of a piece of law being evaluated become about 4.69 times higher if an evaluation clause is present as compared to legislation without such a clause. In terms of predicted probabilities, the chances of an evaluation taking place are 30.9% higher for legislation with an evaluation clause than for legislation without such a clause, if all other variables are kept at their observed values.

Table 2: Results of the logistic regression

	Model 1: Enforcement motive			**Model 2: Strategic motive**		
	B *(SE)*	**Sig.**	**Odds ratio**	**B *(SE)***	**Sig.**	**Odds ratio**
Constant	-2.57 (0.41)	.00	0.08	-2.65 (0.43)	.00	0.07
Legislation type	0.66 (0.29)	.02	1.93	0.72 (0.30)	.02	2.05
Complexity	0.03 (0.01)	.02	1.03	0.03 (0.01)	.02	1.03
Politicization	-0.51 (0.32)	.11	0.60	-0.50 (0.32)	.12	0.60
Council voting				-0.15 (0.36)	.69	0.86
EP involvement				0.46 (0.49)	.35	1.58
Clause	1.57 (0.33)	.00	4.79	1.55 (0.34)	.00	4.69
Unit	0.87 (0.31)	.01	2.38	0.95 (0.32)	.00	2.59
Guidelines	0.73 (0.30)	.01	2.08	0.82 (0.31)	.01	2.26
N	277			277		
Chi^2	62.70			64.62		
Sig	0.00			0.00		
McFadden R^2	0.17			0.17		
AIC	1.184			1.191		

Despite the significance of this variable, it should be noted that only 92 out of 165 pieces of law with an evaluation clause (55.8%) were evaluated, while 24 out of 112 pieces of law without such a clause (21.4%) were evaluated as well. The first number shows that the Commission only complied with a little more than half of the evaluation clauses inserted in major legislation from 2000-2004, indicating that the presence of such clauses is not a guarantee that legislation will be evaluated. The numbers also show that the presence of evaluation clauses only explains a part of the variation in the initiation of EPL evaluations.

Table 2 also shows that both indicators for evaluation capacity are significant. The odds of a piece of law being evaluated are about 2.59 times higher for legislation of a DG with an evaluation unit as compared to legislation of a DG without such a unit, and 2.26 times higher for legislation of a DG that has evaluation guidelines as compared to legislation of a DG without such guidelines. In terms of predicted probabilities, the chances of an evaluation taking place are 18.3% higher in the first case and 15.9% higher in the second case, if all other variables are kept at their observed values.

When interpreting these results, however, it should be noted that high evaluation capacity may be a consequence as well as a cause of evaluation-related activities. For example, it is possible that DGs that initiate more EPL evaluations also invest more in evaluation guidelines to support such evaluative activities. It should also be noted that the results concerning evaluation capacity somewhat depend on our assumptions about the number of years after which it was decided not to evaluate certain legislation (as explained in our methodology section and Note 9). Therefore, more research is needed to establish the exact effect of evaluation capacity on the initiation of EPL evaluations in the EU.

5. Conclusion

This article has sought to describe and explain the variance in the initiation of EPL evaluations by the European Commission. Although the Commission officially endorses EPL evaluations (European Commission, 2007: 3; 2013: 11; 2015: 296), little was known about how systematically the institution conducts such evaluations in practice (but see Mastenbroek et al., 2016). This study aimed to shed light on this underexplored topic by developing a theoretical approach based on two motives for the Commission to evaluate - an enforcement and a

strategic motive - while controlling for other potential explanations. We tested these explanations with the help of binary logistic regression, based on two self-developed datasets.

The results show that less than half of all major EU legislation from 2000-2004 (41.9%) was evaluated. However, the proportion of evaluated legislation has increased over time. Only a small proportion of the major legislation was evaluated more than once.

Concerning the enforcement motive, our results suggest that the odds of being evaluated are significantly higher for directives than for regulations, and that these odds also increase significantly as legislation becomes more complex. This indicates that the Commission prioritizes evaluating legislation for which the chances of non-compliance are relatively high, and that evaluations may at least partly be initiated to scrutinize member state implementation. Concerning the strategic motive, however, we did not find any significant results. This indicates that the risk of EU legislation being significantly amended does not affect its odds of being evaluated.

Two control variables also turned out to be significant. Firstly, the odds of legislation being evaluated increase significantly if that legislation contains an evaluation clause. However, our data also revealed that the Commission only complies with such clauses in about half of all cases. Secondly, the evaluation capacity of the DG that is responsible for the legislation significantly increases the odds of that legislation being evaluated.

In conclusion, our analysis indicates that the initiation of EPL evaluations by the Commission is best explained by a mix of its need to enforce EU legislation towards the member states, its formal obligations to evaluate and its evaluation capacity. However, these conclusions should be viewed in the light of two possible limitations of this research. Firstly, the quantitative nature of our study required us to use indicators that could be measured efficiently for a large number of cases. Some of these indicators may not entirely cover the abstract concepts that they are supposed to represent, such as evaluation capacity and politicization. Therefore, to sustain the conclusions of this article, a follow-up case study with more sophisticated indicators would be useful.

A second limitation of this study is its time period. As explained above, EPL evaluations may be conducted a decade or more after a piece of law enters into force, so we could not yet assess the extent to which legislation from after 2004 has been evaluated without risking a bias

in our data. Whereas our dataset of 277 major regulations and directives (initiated by seventeen DGs) is so broad that our findings are probably not affected by any particular political choice made during 2000-2004, it still seems worthwhile to repeat this research in the future to assess to what extent post-2005 legislation is evaluated.

Two other possibilities for future research stand out. Firstly, since this article showed that the Commission does not always comply with evaluation clauses, a follow-up study about the reasons for this seems worthwhile. Secondly, it could be examined to what extent the factors presented in this study also explain variance in the quality of EPL evaluations, as this is another important characteristic of a proper evaluation system and previous research has shown that the quality of the Commission's EPL evaluations varies greatly (Mastenbroek et al., 2016: 1340-1341).

Notes

1. This refers to all legislation having its legal basis in another regulation or directive.
2. Evaluations published in 2013 and 2014 were added to the dataset used in chapter 2 of this dissertation, resulting in a total of 313 cases. Six further evaluations published in French were excluded because of our lacking language skills. However, including these evaluations would not have changed the results, since they did not study additional legislation from 2000-2004.
3. ec.europa.eu/smart-regulation/evaluation/search/search.do
4. http://ec.europa.eu/atwork/key-documents/index_en.htm
5. https://bookshop.europa.eu/en/home/
6. SWD(2013)228 and SWD(2012)383.
7. http://eur-lex.europa.eu/homepage.html
8. 'evalu*', 'repo*', 'stud*' and 'research'.
9. As described in the methodology section of our article, for legislation that was not evaluated we entered our evaluation capacity data based on the assumption that decisions not to evaluate were taken five years after the publication date of the legislation. We experimented with other time periods than five years for this as well (four, six, seven and eight years). If we assume decisions not to evaluate to be taken after six years instead of five the presence of an evaluation unit is no longer significant, and if we assume this period to be seven years or more both indicators for evaluation capacity seize to be significant. However, assuming a lesser number of years than five does not change the significance of either indicator. The reason for these different findings appears to be that some of the earliest DGs to develop evaluation units and/or guidelines (DG Eurostat, DG Internal Market and DG Employment in particular) also have the highest initiation ratios (as listed in Figure 1). Around 2010 many DGs that have smaller initiation ratios also started to develop such evaluation capacity. Therefore, belonging to a DG that was a frontrunner in building evaluation capacity could be more important than belonging to a DG with high evaluation capacity per se when it comes to explaining variation in the initiation of EPL evaluations.

References

Boswell C (2008) The political functions of expert knowledge: Knowledge and legitimation in the European Union. *Journal of European Public Policy* 15(4): 471-488.

Bovens M, 't Hart P and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford handbook of public policy*. Oxford: University Press, pp. 320-335.

Coglianese C (2012) *Evaluating the performance of regulation and regulatory policy*. Report to the Organization of Economic Cooperation and Development.

Eijlander P and Voermans W (2000) *Wetgevingsleer [Legislative theory]*. The Hague: Boom Juridische Uitgevers.

European Commission (2007) *Communication to the Commission from Ms Grybauskaité in agreement with the President: Responding to strategic needs: Reinforcing the use of evaluation [SEC(2007)213]*. Brussels: European Commission.

European Commission (2010) *Multi-annual overview (2002-2009) of evaluations and impact assessments*. Available at: http://ec.europa.eu/smart-regulation/evaluation/docs/multiannual_overview_en.pdf (Accessed 10 July 2015).

European Commission (2012) *EU regulatory fitness [COM(2012)746]*. Brussels: European Commission.

European Commission (2013) *Regulatory Fitness and Performance (REFIT): Results and next steps [COM(2013)685 final]*. Brussels: European Commission.

European Commission (2015) *Better Regulation Toolbox [SWD(2015)111]*. Brussels: European Commission.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

Franchino F (2000) Control of the Commission's executive functions. *European Union Politics* 1(1): 63-92.

Franchino F (2007) *The powers of the Union: Delegation in the EU.* Cambridge: University Press.

Häge FM (2007) Committee Decision-Making in the Council of the European Union. *European Union Politics* 8(3): 299-328.

Hartlapp M, Metz J and Rauh C (2014) *Which policy for Europe? Power and conflict inside the European Commission*. Oxford: University Press.

Hofmann A (2013) *Strategies of the repeat player. The European Commission between Courtroom and legislator*. PhD Thesis, University of Cologne, Germany.

Kaeding M (2006) Determinants of transposition delay in the European Union. *Journal of Public Policy* 26(3): 229-253.

Kassim H, Peterson J, Bauer MW et al. (2013) *The European Commission of the Twenty-First Century.* Oxford: University Press.

König T (2008) Analysing the Process of EU Legislative Decision-Making: To Make a Long Story Short. *European Union Politics* 9(1): 145-165.

König T and Mäder L (2014) The Strategic Nature of Compliance: An Empirical Evaluation of Law Implementation in the Central Monitoring System of the European Union. *American Journal of Political Science* 58(1): 246-263.

Majone G (1999) The regulatory state and its legitimacy problems. *West European Politics* 22(1): 1-24.

Majone G (2005) *Dilemmas of European integration: The ambiguities and pitfalls of integration by stealth*. Oxford: University Press.

Mastenbroek E (2003) Surviving the deadline: The transposition of EU directives in the Netherlands. *European Union Politics* 4(4): 371-396.

Mastenbroek E, Van Voorst S and Meuwese A (2016) Closing the regulatory cycle? A meta-evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy* 23(9): 1329-1348.

McCormick J (2015) *European Union Politics (2nd edition)*. London: Palgrave.

Nielsen SB, Lemire S and Skov M (2011) Measuring evaluation capacity: Results and implications of a Danish study. *American Journal of Evaluation* 32(3): 324-344.

OECD (2015) *OECD Regulatory Policy Outlook 2015*. Paris: OECD Press.

Pattyn V (2014) Why organizations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation* 20(3): 348-367.

Pollack MA (2008) M*ember-State Principals, Supranational Agents, and the EU Budgetary*

Process, 1970-2008. Paper presented at the Conference on Public Finances in the European Union in Brussels, 3-4 April 2008.

Poptcheva EM (2013) *Library Briefing. Policy and legislative evaluation in the EU*. Brussels: European Parliament.

Radaelli CM (1999) The public policy of the European Union: Whither politics of expertise? *Journal of European Public Policy* 6(5): 757-774.

Radaelli CM and Meuwese ACM (2010) Hard questions, hard solutions: Proceduralisation through impact assessment in the EU. *West European Politics* 33(1): 136-153.

Schmidt SK and Wonka A (2013) The European Commission. In: Jones E, Menon A and Weatherill S (eds) *The Oxford Handbook of the European Union*. Oxford: University Press, pp. 336-349.

Schwartz R (1998) The Politics of Evaluation Reconsidered: A Comparative Study of Israeli Programs. *Evaluation* 4(3): 294-309.

Smith M (2015) Evaluation and the Salience of Infringement Data. *European journal of risk regulation* 6(1): 90-100.

Stame N (2008) The European project, federalism and evaluation. *Evaluation* 14(2): 117-140.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 67-85.

Steunenberg B and Rhinard M (2010) The Transposition of European Law in EU Member States: Between Process and Politics. *European Political Science Review* 2(3): 495-520.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation*. New Brunswick: Transaction Publishers, pp. 407-424.

Torriti J (2010) Impact assessment and the liberalization of the EU energy markets: Evidence-based policy-making or policy-based evidence-making? *Journal of Common Market Studies* 48(4): 1065-1081.

Treib O (2014) Implementing and complying with EU governance outputs. *Living Reviews in European Governance* 9(1): 1-47.

Van Golen T and Van Voorst S (2016) Towards a regulatory cycle? The use of evaluative information in impact assessments and ex-post evaluations in the European Union. *European Journal of Risk Regulation* 7(2): 388-403.

Van Voorst S (2017) Evaluation capacity in the European Commission. *Evaluation* 23(1): 24-41.

Vedung E (1997) *Public policy and program evaluation.* New Brunswick: Transaction.

Weiss CH (1993) Where Politics and Evaluation Research Meet. *American Journal of Evaluation* 14(1): 93-106.

Wille A (2010) Political-bureaucratic accountability in the EU Commission: Modernising the executive. *West European Politics* 33(5): 1093-1116.

Wille A (2013) *The normalization of the European Commission: Politics and bureaucracy in the EU executive*. Oxford: University Press.

Zhelyazkova A, Kaya C and Schrama R (2016) Decoupling practical and legal compliance: analysis of member states' implementation of EU policy. *European Journal of Political Research* 55(4): 827-846.

Zwaan P, Van Voorst S and Mastenbroek E (2016) Ex-post regulatory evaluations in the European Union: questioning the use of evaluations as instruments for accountability. *International review of administrative sciences* 82(4): 674-693.

Chapter 5: The quality of the European Commission's ex-post legislative evaluations

Stijn van Voorst and Ellen Mastenbroek

Paper not yet published.

Abstract

Ex-post legislative (EPL) evaluations are a potentially important tool for the European Commission to learn how EU legislation can be improved. To prevent such learning from taking place based on false or incomplete information, the evaluations must be of sufficient methodological quality. This paper therefore describes and explains the variance in the quality of the European Commission's EPL evaluations. A number of potential political and technical explanations for this variance are tested with a self-constructed dataset of 153 EPL evaluations. The results show that the Commission's EPL evaluations do well in terms of applying a robust methodology, but that the clarity of their scope, the accuracy of their data and the foundations of their conclusions are problematic. The variance in this quality is mainly explained by the type of evaluator: EPL evaluations conducted by external actors are of higher quality than evaluations conducted internally by the Commission.

1. Introduction

The European Union (EU) is frequently called a 'regulatory state' because of its reliance on legislation as its main policy instrument (Majone, 1999: 1; Scharpf, 1999: 189). Although the EU's legislative output has decreased in recent years, it still produces dozens of regulations and directives per year (European Commission, 2016: 2-3). On paper, this legislation is supposed to solve a wide range of economic and social problems.

To improve the effectiveness of EU legislation in solving such problems, the European Commission (2007: 3; 2013: 6; 2015: 7; 2016: 2) has developed a 'better regulation' programme since 2000. The first key element of this programme was the introduction of impact assessments (IAs): forward-looking evaluations aimed to increase the quality of legislative proposals (European Commission, 2007: 4). Initially, however, the effectiveness of legislation in practice received less attention. This changed in 2007, when the Commission (2007: 3; 2015: 253) pledged to further develop its system for ex-post legislative (EPL) evaluation: reports that retrospectively assess the effects of EU legislation.

EPL evaluations are supposed to produce evidence that allows the Commission to learn whether or not legislation achieves its intended goals and how its impact can be improved (Fitzpatrick, 2012: 479; Vedung, 1997: 109). Theoretically, this learning function is important because the policies created by modern-day governments are so complex that their effects on society are inherently uncertain (Bunea and Ibenskas, 2017: 592; Sanderson, 2002: 1, 7). In order to make sound decisions about which instrument is best suited to achieve certain objectives at a given moment in time, governments require systematic and continuous evidence about the working of their policies (Böhme, 2002: 99; Bunea and Ibenskas, 2017: 592; Sanderson, 2002: 3, 5). Arguably this requirement is especially pertinent for EU legislation, which has highly uncertain effects because it needs to be implemented in almost thirty member states with very different policy traditions (Fitzpatrick, 2012: 480-481).

Theoretically, the Commission's EPL evaluations may contribute to evidence-based policy making in at least four ways. First, the evidence produced by EPL evaluations is supposed to feed into the IAs that the Commission attaches to legislative proposals (European Commission, 2015: 258; 2016: 3). In particular, IAs can use findings from EPL evaluations to compare the effectiveness of existing legislation to the costs and benefits of alternative policies. Research has shown that about 65% of the Commission's IAs use results from EPL evaluations when they are available, which suggests that this regulatory cycle functions in practice to some extent (Van Golen and Van Voorst, 2015: 402). Second, evidence provided by EPL evaluations can be used by the Commission to improve delegated acts or to develop guidelines for the practical implementation of legislation. Third, EPL evaluations can provide

information to stakeholders about a particular piece of legislation and allow them to push for changes accordingly (European Commission, 2015: 280). Fourth, EPL evaluations can allow national governments to learn what is the best way to implement EU legislation.

However, EPL evaluations will not automatically lead to learning. This link is contingent on a key necessary condition, central to this paper: high evaluation quality (Forss and Carlsson, 1997: 481). Evaluation quality is a key precondition for learning for two reasons. Firstly, the information provided by EPL evaluations only enhances learning if it is accurate and clear. When actors try to learn what policies work based on evaluations containing false or misleading data, the decisions that they make are unsubstantiated, which may lead to poor decision-making (Cooksy and Caracelli, 2005: 31; Mayne and Schwartz, 2005: 7; Sanderson, 2002: 13). Secondly, since evaluations claim to describe reality objectively, their credibility is thwarted if actors find out that they contain misleading information. When this happens, decision-makers are likely to distrust further evaluations and the potential of such reports for learning is lost (Cooksy and Caracelli, 2005: 31; Mayne and Schwartz, 2005: 7). Therefore, EPL evaluations only enable policy learning if they observe certain standards of methodological quality.

The Commission's system for EPL evaluation is designed to enhance quality in three main ways. Firstly, whereas EPL evaluations are primarily the responsibility of the Commission's Directorates-General (DGs), its Secretariat-General (SG) may set quality standards that all DGs must observe (European Commission, 2007: 22-24; 2015: 252-298). Secondly, DGs outsource most of their EPL evaluations to specialized consultants to boost their technical quality (European Commission, 2015: 282-9; Van Voorst, 2017: 33-34). In such cases the consultants conduct most of the evaluation, while the responsible DG monitors the quality of their work (European Commission, 2015: 337-414; Fitzpatrick, 2012: 490-7). Thirdly, the Commission's Regulatory Scrutiny Board (RSB) (2018: 11) annually judges the quality of a small number of EPL evaluations. These reports are often revised when the RSB's (2018: 12) opinion is negative.

Despite the theoretical importance of the topic, empirical research about the quality of the Commission's EPL evaluations has so far been limited. Whereas both academics (Cecot et al., 2008: 412-6; Lee and Kirkpatrick, 2004: 17-20; Renda, 2006: 62-66) and the RSB of the Commission (2018: 11) have shown that the quality of impact assessments is often below

standard, similar research about the quality of the Commission's EPL evaluations is less systematic. While an earlier explorative study (Mastenbroek et al., 2016: 1340) and the reports from the RSB (2018: 11) showed the quality of the Commission's EPL evaluations to vary considerably, a more comprehensive and explanatory analysis of the topic is missing. This paper aims to fill this gap by answering the following research question: *how can the variance in the quality of ex-post legislative evaluations by the European Commission be explained?*

In answering this question, we consider two sets of factors that may compromise the quality of EPL evaluations and, hence, their potential to inform learning. A first potential threat is political influence. Evaluation results can be used strategically in the political arena, for example to shift blame between actors or to criticize policies (Bovens et al., 2008: 319; Versluis et al., 2011: 213-214; Weiss, 1993: 94). The criticism that results from repeated negative evaluations can threaten the legitimacy of public organizations (Weiss, 1993: 95). This risk may be especially threatening for the Commission, as its unelected nature and the existence of Euroscepticism cause its activities to be constantly scrutinized (Versluis et al., 2011: 207). The Commission may therefore have an incentive not to learn to what extent its legislation is effective, especially when it expects the results of such assessments could be negative and hence open up discussion about its own role.

One way to deal with unwelcome evaluations is to not initiate them in situations where negative outcomes are likely (Van Voorst and Mastenbroek, 2017: 645). However, in some cases this may be impossible, for example because an EPL evaluation is made compulsory by legal requirements. When an evaluation with potentially negative outcomes must be conducted, institutions like the Commission may have an incentive to manipulate their content, for example by selecting evaluation questions and methods that make convenient outcomes more likely (Chelimsky, 2008: 404; House, 2008: 418). In such cases, evaluation quality is compromised.

Secondly, the quality of the Commission's EPL evaluations may be thwarted by 'technical' factors, inherent to the methodological difficulties of evaluating legislation. A first technical explanation is evaluation capacity: DGs that have more/better resources and procedures in place regarding evaluations may be able to produce better reports, as such

resources and procedures allow for extra investments in every phase of the evaluation process (Forss and Carlsson 1997: 498; Nielsen et al., 2011: 325-327; Rossi et al., 2004: 414).[1] The type of evaluator could also matter: external parties may produce evaluations of higher quality than internal ones, given the fact that they have better expertise (Vedung, 1997: 117). A third technical explanation is complexity: for some legislation it is more difficult to produce a high-quality evaluation than for other legislation (Bussmann, 2010: 281).

The hypotheses flowing from these two types of explanations are tested using a dataset of 153 EPL evaluations conducted or outsourced by the Commission during 2000-2014. The results show that the Commission's EPL evaluations do well in terms of applying a robust methodology, but that the clarity of their scope, the accuracy of their data and the foundations of their conclusions are problematic. The variance in this quality is mainly explained by the type of evaluator: EPL evaluations conducted by external actors are of higher quality than evaluations conducted internally by the Commission.

2. Conceptualizing evaluation quality

This paper uses four methodological criteria to assess the quality of the Commission's EPL evaluations. These criteria have been derived from Mayne and Schwartz (2005: 304-305), who developed them based on the standards used by countries and international organizations that are frontrunners in the field of policy evaluation. These standards in turn are often based on general methodological criteria from the social sciences, as evaluations are essentially a form of applied social research (Mayne and Schwartz, 2005: 305).

The first criterion is a *well-defined scope*: the purpose and the topic of an evaluation must be properly specified (Mayne and Schwartz, 2005: 304). If this criterion is not met, it remains unclear what an evaluation's results are about and how broadly they can be applied. These issues in turn make it difficult to learn from an evaluation. In the context of this paper, the criterion of a well-defined scope means that the Commission's EPL evaluations should clearly specify which intended outcomes of which piece of legislation they study.

The second criterion is *accurate data*: the raw information presented by an evaluation must be valid and reliable (Mayne and Schwartz, 2005: 305). Validity refers to the absence of systematic errors in research results; reliability concerns the absence of random errors (Adcock and Collier, 2001: 531). Validity can be further split into two types. The first type is internal or *content validity*: the correct measurement of abstract concepts (Adcock and Collier, 2001: 538). The second type is *external validity*: the degree to which results based on a sample represent a whole population (Adcock and Collier, 2001: 529).

In the context of this paper, validity and reliability mean that the Commission's EPL evaluations should avoid both systematic and random errors when assessing the effectiveness of legislation. If this is not the case, learning about this effectiveness occurs based on false information. This in turn may lead to poor or unsubstantiated decision-making (Cooksy and Caracelli, 2005: 31; Mayne and Schwartz, 2005: 7). External validity also matters for the Commission's EPL evaluations because they must often make some selection of member states or stakeholders (Fitzpatrick, 2012: 490). For such evaluations to contribute to learning about the effectiveness of an entire piece of legislation, the results for the selected countries or actors must correctly represent the situation in the whole EU.

The third criterion is *robust methodology*:[2] evaluations should use methods that fit their research objective (Mayne and Schwartz, 2005: 305). In the context of EPL evaluations experimental methods are often impossible, as legislation is universal and therefore leaves no room for a control group (Bussmann, 2010: 281; Coglianese, 2012: 404). Conversely, methodologies that involve stakeholders are highly fitting when evaluating EU legislation, as there are many different actors involved in the implementation of such policies (like member states, local governments and interest groups) (Fitzpatrick, 2012: 481, 489). Stakeholders who implement policies in their daily work presumably have the best view on how they function in reality (Varvasovszky and Brugha, 2000). Therefore, considering the views of a multitude of stakeholders is essential to learn if and why EU legislation works.

The fourth criterion is *substantiated findings*: an evaluation's conclusions should be based on its underlying data (Mayne and Schwartz, 2005: 305). This criterion matters because actors who seek to use an evaluation (for learning or other aims) may not have time to read it

entirely and may therefore rely on its conclusions only (Vedung, 1997: 281). When such conclusions are not clearly related to the underlying data, decision-makers may have insufficient details to fully learn how a policy works (Coglianese, 2012: 62-63). Furthermore, if a conclusion is not clearly supported by underlying data, this may create distrust in the validity of evaluations' findings, making them less useful for learning (Forss and Carlsson, 1997: 481, 490).

3. Theoretical framework

Although much academic literature about policy evaluation exists, there is no comprehensive theory that explains variation in evaluation quality (Mastenbroek et al., 2016: 1343). Therefore, this paper develops an explanatory model for evaluation quality based on broader theories about EU governance and policy evaluation. In this model we considers two types of variables that may affect evaluation quality: political and technical explanations.

Political explanations

EPL evaluations can be perceived as strategic tools in the hands of decision makers. This idea is rooted in the theoretical view that evaluations are never entirely neutral: their results are always advantageous to some actors while being disadvantageous to others (Bovens et al., 2008: 319; Chelimsky, 2008: 400; Versluis et al., 2011: 213-214; Weiss, 1993: 95-96). For example, evaluations can be used strategically to delay decisions, to shift responsibilities for mistakes and to provide a semblance of rationality (Vedung, 1997: 111-13).

Political pressure is generally considered a threat to evaluation quality (Cooksy and Mark, 2012: 82; Datta, 2011: 281). Potentially, it has a negative effect on all four quality criteria described above (Mayne and Schwartz, 2005: 314-316). Political influence may prevent a *well-defined scope* when an evaluation's client prescribes vague or suggestive research questions to serve his own interests (Chelimsky, 2008: 404; Vedung, 1997: 93-94). *Accurate data* is difficult to collect when actors refuse to provide evaluators with information that could harm their political position (Chelimsky, 2008: 404; Weiss, 1993: 96) or this information gets distorted because stakeholders are selectively involved in the evaluation process (Bovens et al., 2008:

321; House, 2008: 417). The criterion of *robust methodology* is not met when unwelcome parts of data are ignored (Chelimsky, 2008: 401; House, 2008: 418). Finally, political pressure may have a negative effect on *substantiated findings* when the conclusion of an evaluation is rewritten to include findings favourable to specific actors or to drop results that are unwelcome to them (Chelimsky, 2008: 404; House, 2008: 418).

Previous research has shown that the Commission deals strategically with the initiation and use of both impact assessments (Poptcheva, 2013: 4; Torriti, 2010: 1065) and EPL evaluations (Van Voorst and Mastenbroek, 2017: 653; Zwaan et al., 2016: 688). Therefore, we can expect strategic considerations to influence the quality of the Commission's EPL evaluations as well. Below this general expectation is translated into a number of specific hypotheses.

Arguably, the risk of strategic considerations affecting evaluation quality is especially high when an evaluation's topic is politically controversial (Boswell, 2008: 473-476). According to this logic, the sensitivity of a piece of legislation increases the stakes of the actors involved in the evaluation process, which in turn increases the chances that they will take note of the evaluation and will attempt to distort its results in some of the ways that were described above. This may reduce the quality of the evaluation.

Hypothesis 1: The more politicized the topic of an EPL evaluation, the lower the quality of that evaluation.

A second important strategic consideration is that evaluations with negative results can lead to demands to reduce the role of actors responsible for policy implementation (Vedung, 1997: 102-108). In the context of the EU, EPL evaluations may be used by the European Parliament (EP), the Council and other stakeholders to scrutinize the Commission's activities (Radaelli and Meuwese, 2010: 138; Versluis et al., 2011: 208; Zwaan et al., 2016: 688). Therefore, EPL evaluations with negative findings may lead such actors to call for the Commission's competences to be reduced and/or for policies to be 'repatriated' to the national level. We thus expect that the Commission will particularly wish to influence the results of EPL evaluations when there is a risk that these results may lead to significant legislative amendments.

Involvement of the EP in decision making decreases the chances of significant amendments and is therefore expected to have a positive effect on evaluation quality. The reason for this is that the EP provides an extra veto player that can block amendments (Häge, 2007: 307). As a majority of EP members generally supports further European integration (Pollack, 2008: 9), it can also be expected that the EP will usually oppose reducing the competences of supranational institutions like the Commission.

Hypothesis 2: Evaluations of pieces of legislation for which the European Parliament has veto powers are of higher quality than evaluations of pieces of legislation for which the European Parliament does not have veto powers.

The *voting procedure in the Council* is also expected to influence the chances of legislative amendments. If unanimity is required in the Council, it is significantly harder to change legislation, as it is difficult to make all member states agree on a proposal (Häge, 2007: 308). Theoretically, this difficulty to amend legislation reduces the risk that negative evaluation results will threaten the Commission's competences and therefore decreases the incentive for the Commission to distort evaluation results. We therefore expect the quality of an evaluation to be higher when the Council applies unanimity voting, as compared to when it applies qualified majority voting (QMV).

Hypothesis 3: Evaluations of pieces of legislation decided upon by unanimity in the Council are of higher quality than evaluations of pieces of legislation decided upon by qualified majority voting.

The quality of the Commission's EPL evaluations may also be affected by the presence of *evaluation clauses* in the evaluated legislation. Such clauses include legal obligations to evaluate the legislation in a certain way and at a certain moment in time (Summa and Toulemonde, 2002: 410). These legal obligations may cause EPL evaluations to become 'tick-the-box exercises' that are only conducted because they are obligatory rather than as genuine

efforts to learn about policies (Cooksy and Mark, 2012: 82; Radaelli and Meuwese, 2010: 146). When there is a lack of enthusiasm to evaluate, the quality of EPL evaluations may suffer. Evaluation clauses also prevent flexibility, as the timeframe of three to five years that they tend to prescribe may be too short to conduct a proper EPL evaluation.[3]

Hypothesis 4: Evaluations of pieces of legislation containing an evaluation clause are of lower quality than evaluations of pieces of legislation containing no evaluation clause.

Technical explanations

Besides the political variables described above, evaluation quality may also be affected by 'technical' factors. Such factors are rooted in an apolitical or rationalistic view on policy evaluation. This perspective encompasses the idea that evaluations can produce objective knowledge when the correct procedures are followed and the right evaluators are involved in the process, regardless of political context (Bovens et al., 2008: 325).

Three specific technical factors may affect the quality of the Commission's EPL evaluations. The first is *evaluation capacity*: the means and procedures meant to ensure that high-quality evaluations are ongoing practices within organizations (Nielsen et al., 2011: 325). Higher evaluation capacity can be expected to positively affect evaluation quality (Cooksy and Mark, 2012: 81), as it allows for more investments in every stage of an evaluation process. For example, having a staff that is trained well in evaluation methods can lead to better data collection and analysis (accurate data and robust methodology) (Nielsen et al., 2011: 327). Furthermore, more evaluation capacity allows for extra investments in writing high-quality reports (Forss and Carlsson 1997: 498; Rossi et al., 2004: 414), which could lead to a more thorough description of the evaluated policy (well-defined scope) and to results being presented in a way that clearly links them to the underlying data (substantiated findings).

Within the Commission, the DGs are the main organizational units that conduct evaluations (Stern, 2009: 71). Existing research (Van Voorst, 2017: 33) shows that the capacity of these DGs to conduct EPL evaluations varies greatly: some DGs have clear procedures for EPL evaluations in place and invest much financial and human capital in them, while for other DGs

less capacity is available. We expect that these capacity differences between DGs (partly) explain the variance in the quality of the Commission's EPL evaluations.

Hypothesis 5: The higher the evaluation capacity of the DG that conducts an evaluation, the higher the quality of that evaluation.

A second technical factor that may affect evaluation quality is the *type of evaluator*. Although the Commission's DGs outsource most of their EPL evaluations, they may also conduct them internally (European Commission, 2015: 282-289). External evaluations can be expected to be of higher quality than internal ones, as they are generally conducted by more experienced evaluators (Vedung, 1997: 117). Most DGs sign multi-annual framework contracts with specialized consultants, which allows these companies to gain expertise in evaluating EU legislation (Van Voorst, 2017: 33-34). This expertise may result in improvements to all elements of evaluation quality, including the presence of a well-defined scope, accurate data, robust methodology and substantiated findings. Although DGs must also employ some evaluation experts internally, this is usually a small coordinating staff that has less experience with conducting full evaluations (Van Voorst, 2017: 33).

Some academics (e.g. Vedung, 1997: 117) suggest that external evaluators also produce better reports because they are more impartial than internal evaluators. Other literature disputes this claim, as the fact that consultants depend on policy makers for their future funding may give them an incentive not to be too critical (Conley-Tyler, 2005: 7). On the other hand, if external evaluators are too sensitive to their clients' interests, this may work as a 'boomerang', by affecting a core building block of their reputation: producing objective reports. All in all, we do not expect the difference in impartiality to affect the quality of EPL evaluations in the context of this paper.

Hypothesis 6: External EPL evaluations are of higher quality than internal ones.

The third technical factor that may affect quality is *legislative complexity.* Legislation is often difficult to evaluate because it contains multiple overlapping interventions with different goals (Bussmann, 2010: 281). This is especially the case in the context of the EU, where regulations and directives are based on extensive compromises between different member states and supranational institutions (Delahais, 2014: 7, 9; Fitzpatrick, 2012: 480-481; Häge, 2007: 307-308). The implementation of EU legislation also typically involves a complex web of actors (Delahais, 2014: 9; Fitzpatrick, 2012: 480; Steunenberg, 2006: 294-295).

Such complexity can make it difficult to conduct high-quality EPL evaluations (Bussmann, 2010: 281; Delahais, 2014: 1; Fitzpatrick, 2012: 480-481). In particular, the fact that EU legislation often contains multiple goals and interventions can make it challenging to clearly delineate the topic of an EPL evaluation (well-defined scope). For complex legislation it may also be more difficult to identify and gain access to all stakeholders and to find other appropriate sources of information (accurate data and robust methodology) (Bussmann, 2010: 281; Fitzpatrick, 2012: 480-481). Furthermore, for more complex legislation it may be harder to define when it should be considered successful, which makes it more difficult to draw conclusions about its effectiveness from the available evidence (substantiated findings).

Hypothesis 7: The more complex the piece of legislation that is evaluated, the lower the quality of the evaluation.

4. Methods and data

Data collection

This research uses a self-constructed dataset of 153 evaluations in which the Commission or an external party hired by the Commission retrospectively assessed the effectiveness of European regulations or directives. Evaluations focusing on process aspects or side effects of legislation only were left out due to this paper's focus on learning about effectiveness. The dataset starts at January 2000 because of the lacking online availability of earlier EPL evaluations; it ends at

December 2014 because not all evaluations from 2015 and later had been published online when the data collection was completed. Three other types of EPL evaluations were discarded. Firstly, evaluations of legislation that only regulates the EU institutions or actors outside of the EU were left out, as the Commission's better regulation agenda focuses on legislation that affects citizens and companies (European Commission, 2007: 3; 2016: 2). Secondly, we discarded all evaluation reports that merely summarize other evaluation reports. Thirdly, four EPL evaluations only available in French were left out because reading them would have required extensive knowledge of that language.

The evaluations were collected from various sources: the Commission's search engine for evaluations,[4] the Commission's multi-annual evaluation overview (2010), EU bookshop,[5] annexes to the Commission's financial reports,[6] the Commission's work programmes,[7] and lists of evaluations found on websites of the Commission's DGs. The data collection was checked by using an existing dataset of evaluations from expertise centre Eureval, by running Google searches for evaluations of major legislation adopted between 1996 and 2010, by searching for background documents of legislation in Eur-lex,[8] and by discussions with the SG. For a further description of the dataset, see Mastenbroek et al. (2016: 1334-1335) (chapter 2 of this dissertation).

Operationalization of evaluation quality

The quality of each evaluation report was measured by coding it with the help of a standardized scorecard. This method has the advantage that it allows for studying a large number of evaluations in a short amount of time (Forss and Carlsson, 1997: 483). Its disadvantage is that it is unfit to judge evaluation processes, which are usually not described in the reports. The scorecard method also does not allow for in-depth judgements of the content of individual evaluations. Therefore, this paper focuses on characteristics of the reports that can be efficiently measured.

The criterion of a *well-defined scope*, firstly, was measured using two indicators. The first indicator is the presence of a clear problem definition: the report should mention its aim to measure the effectiveness of specific legislation before presenting its findings. The second

indicator is the presence of a reconstruction of the legislation's intervention logic: an overview of the steps through which the regulation or directive was intended to reach its goals. Such reconstructions matter because evaluations that seek to understand a policy's effectiveness should first map how it was meant to work (Fitzpatrick, 2012: 485; Stern, 2009: 70).

Secondly, concerning *accurate data*, to check if an EPL evaluation measures effectiveness without too many errors, the various types of validity and reliability discussed in the theoretical section were measured. Content validity was assessed by checking the evaluations for the presence of a clear operationalization: a list of concrete indicators used to measure effectiveness. External validity was measured by using two indicators: a representative selection of member states and a representative selection of cases within these states. Unless all countries or cases were selected, the evaluation had to provide a clear explanation for the representativeness of its selection. Reliability was assessed by checking the replicability of the evaluations: do the reports provide their questionnaires, lists of respondents and the like, so that the research could be repeated?

The criterion of a *robust methodology*, thirdly, was assessed by checking the evaluations for the presentation of at least some stakeholder opinions regarding legislative effectiveness. A second indicator was the use of triangulation: are the evaluation's findings based on at least two different methods of data collection? Triangulation is a sign of methodological robustness because it allows for double-checking findings about effectiveness. The following methods were counted as substantially different when measuring triangulation: studying existing content, direct observations, surveys, focus groups and interviews.

The criterion of *substantiated findings*, fourthly, was assessed by checking if the reports clearly link their conclusions to their results. Specifically, the evaluations were required to (1) contain a conclusion that judges the legislation's effectiveness and (2) provide clear sources or references to data presented earlier in the report in a majority of this conclusion's paragraphs. Only paragraphs that answered research questions were included in this calculation: opening paragraphs and paragraphs that merely served to structure the conclusion were not counted.

Each indicator presented above was measured dichotomously, with evaluation reports that provided no information about a certain indicator automatically being coded as zero.

Twenty cases were coded by both authors of this paper to assess intercoder reliability, which was found to be sufficient.[9] Table 1 summarizes the indicators of the scorecard.

Table 1: Operationalization of evaluation quality

Quality criterion	**Indicator(s)**
Well-defined scope	Clear problem definition Intervention logic reconstructed
Accurate data	Content validity: clear operationalization External validity: representative country selection External validity: representative case selection Reliability: replicability
Robust methodology	Stakeholder consultation Triangulation
Substantiated findings	Substantiated conclusions

Operationalization of independent variables

Politicization (hypothesis 1) was measured by establishing whether or not the evaluated legislation was on the Council's agenda as a B-point, as B-points are handled at the political/ministerial level (Häge, 2007: 303). *Involvement of the European Parliament* (hypothesis 2) was measured by assessing the formal procedure used to enact the evaluated legislation as stated by Eur-lex.7 In case of the ordinary legislative procedure (former codecision and cooperation procedures) this involvement was considered high, while in case of the consultation or comitology procedures it was considered low (Häge, 2007: 316). The *voting procedure in the Council* (hypothesis 3) was also measured dichotomously (QMV or unanimity) using Eur-lex. The *presence of an evaluation clause* (hypothesis 4) was measured dichotomously (yes/no) by searching each evaluated piece of legislation using specific keywords.[10]

Evaluation capacity (hypothesis 5) was measured using interviews with the evaluation coordinators of seventeen DGs of the Commission (Van Voorst, 2017: 29-31). Data were collected about twelve capacity indicators, but out of these only the presence of a dedicated (sub)unit for evaluations (yes/no) and the presence of evaluation guidelines (yes/no) could be established per DG per year and were therefore useful for this research. The *type of evaluator*

(hypothesis 6), which could be either internal or external, was deduced from the title page of each report. The *complexity* of the evaluated legislation (hypothesis 7) was measured by its number of recitals, as a larger number of explanations is generally required for more complex legislation (Kaeding, 2006: 236; Steunenberg and Rhinard, 2010: 501).

Table 2: Operationalization of independent variables

Type of explanation	**Variable and hypothesis number**	**Indicator**	**Descriptive statistics**
Political	Politicization Council (H1)	0 = not discussed as B-point 1 = discussed as B-point	0 = 64 cases 1 = 88 cases
	EP involvement (H2)	0 = consultation procedure 1 = ordinary legislative procedure	0 = 40 cases 1 = 112 cases
	Council voting procedure (H3)	0 = QMV 1 = unanimity	0 = 129 cases 1 = 20 cases
	Evaluation clause (H4)	0 = no evaluation clause 1 = evaluation clause present	0 = 39 cases 1 = 113 cases
Technical	Evaluation capacity (H5)	0 = no evaluation unit 1 = evaluation unit present	0 = 88 cases 1 = 65 cases
		0 = no evaluation guidelines 1 = evaluation guidelines present	0 = 75 cases 1 = 78 cases
	Type of evaluator (H6)	0 = internal 1 = external	0 = 33 cases 1 = 120 cases
	Complexity (H7)	Number of recitals	Mean = 28.8 σ = 23.7

Some of the indicators presented above are derived from the evaluated legislation. When multiple pieces of legislation were studied by one evaluation, the average score for these pieces of legislation was used to code continuous variables and the type of the majority of the pieces of legislation was used to code the categorical ones. For example, if an evaluation studied two pieces of legislation with an evaluation clause and one piece of legislation without such a clause, it was coded as containing a clause. In the rare case of a tie, we used the value of the most recent legislation. Table 2 summarizes the operationalization, which has partly been derived from chapter 4 of this dissertation.

Method of analysis

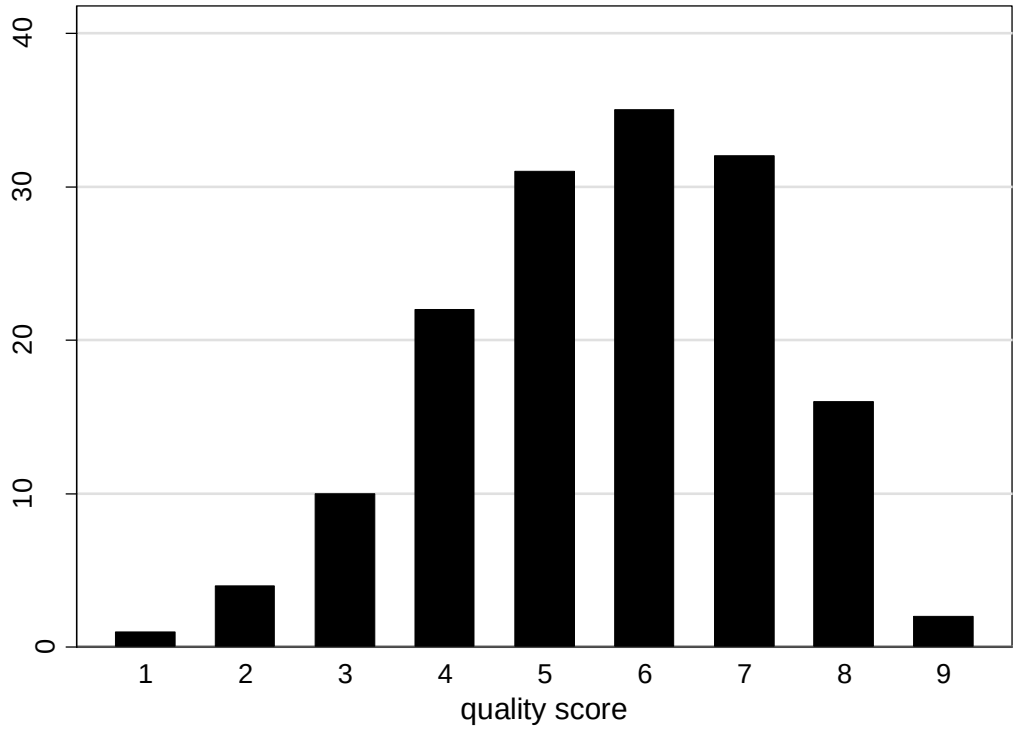
The data presented above were analysed using Ordinary Least Squares (OLS) regression. One assumption of this technique is that the dependent variable is continuous (Miles and Shevlin, 2001: 62), which is not the case for evaluation quality as measured in this paper. However, if the number of categories used to measure an ordinal variable is large enough (e.g. higher than seven) it can still be analysed with OLS regression if all other assumptions of this method are met (Miles and Shevlin, 2001: 62). We carefully checked for these assumptions and found that none of them were violated by our data. We preferred OLS regression over regression methods tailored towards ordinal variables because its results are easier to interpret. The variables were entered in two blocks that match the two types of explanations described above.

5. Results

Descriptive analysis

Figure 1 depicts the variance in the quality of the 153 evaluations studied in this paper. The average quality score is 5.6 on a nine-point scale; about 75% of the reports meet the majority (five or more) of the criteria. No reports received a score of zero and only five reports received a score of one or two. At the other end of the spectrum, two reports[11] meet all of the criteria and sixteen reports meet all but one of them.

Figure 1: distribution of evaluation quality



Overall, this assessment of the 153 cases that focus on effectiveness provides a more positive picture than an earlier study of 216 EPL evaluations produced by the Commission between 2000 and 2012, where the average quality score was 4.1 on an eight-point scale, 43% of the reports met five or more out of eight criteria and two reports received a score of zero (Mastenbroek et al., 2016: 1340). This comparison suggests that the Commission's EPL evaluations that assess effectiveness are of a somewhat higher quality than its ex-post evaluations of other aspects of legislation (like transposition, implementation or side effects).

Table 3 presents the number of EPL evaluations with a positive score per aspect of quality. The table shows that there are large differences between the criteria. The vast majority of the evaluations apply stakeholder consultation and at least one other data collection method (about 76% has a positive score regarding both stakeholder consultation and triangulation), which means that the criterion of a robust methodology is generally being met. The criterion of a well-defined scope is only partly met: although almost all of the Commission's EPL evaluations

include a clear problem definition (89%), only a minority of them goes beyond that by also presenting the intervention logic through which the evaluated legislation is supposed to achieve its aims (37%). Substantiated conclusions are present in a small majority of 57% of the reports, which means that more than four out of ten evaluations have no conclusion that can be clearly linked to its collected data.

Table 3: scores for individual quality aspects from high to low

Indicator	**Quality Criterion**	**Number of reports with positive score (out of 153)**	**%**
Triangulation	Robust methodology	137	90
Clear problem definition	Well-defined scope	136	89
Stakeholder consultation	Robust methodology	128	84
Representative country selection	Accurate data	107	70
Clear operationalization	Accurate data	98	64
Substantiated conclusions	Substantiated findings	87	57
Representative case selection	Accurate data	65	42
Intervention logic reconstructed	Well-defined scope	56	37
Replicability	Accurate data	48	31

Overall, the criterion of *accurate data* is met by the smallest proportion of EPL evaluations. Although 70% of the reports study all member states or clearly explain their selection of certain countries, only 42% of them are fully transparent about how they selected cases within these states. This shows that the external validity of many EPL evaluations is questionable: can their findings be used to learn about the effectiveness of the legislation in general or do they only apply to the specific cases that were studied? Some 64% of the EPL evaluations present a clear operationalization that shows how the legislation's effectiveness was measured, which is important for their internal validity. However, few evaluations meets the standard set for reliability: only 31% of them present all the information that would be needed to repeat their

underlying research. In particular, many of the EPL evaluations present either their interview guides, their questionnaires or their lists of respondents, but not all of this information together, which makes it impossible to check the data collection if required.

Explanatory analysis

Table 4 presents the results of the regression analysis. As the table shows, the model with the political factors only (model 1) is significant at the 0.05 level, but explains just 6% of the variance in the quality of the EPL evaluations. Furthermore, none of the individual independent variables included in this model turn out to be significant in the way the hypotheses predicted. The level of politicization of the evaluated legislation (hypothesis 1), the procedure through which it was enacted by the Council (hypothesis 3) and the presence of an evaluation clause in its text (hypothesis 4) do not explain the variance in the quality of subsequent EPL evaluations.

The voting procedure used in the EP does provide a significant explanation, but its effect is the opposite of what we expected based on our theoretical framework (hypothesis 2). On average, the quality of EPL evaluations of legislation enacted through the ordinary legislative procedure is about one point *lower* than the quality of EPL evaluations of legislation enacted through the consultation procedure. This effect remains significant no matter which of the other factors are included in the model. This result suggests that, contrary to our expectations, the Commission's interest to protect its competences does not explain the variation in the quality of its EPL evaluations. However, it should be noted that the indicators used to measure this interest were fairly general proxies.

One possible reason for the fact that EPL evaluations of legislation enacted through the ordinary procedure are of relatively low quality could be that such legislation is more closely scrutinized by the EP than legislation enacted through the consultation procedure (Rasmussen and Toshkov, 2010: 92). The Commission could therefore have an incentive not to provide the EP with EPL evaluations that can be used for the purpose of this scrutiny. However, more (qualitative) research about the mechanisms behind the Commission's EPL evaluations would be needed to assess the plausibility of this hypothesis.

Table 4 shows that when the technical variables are added (model 2), the model as a

whole is still significant at the 0.05 level and its explanatory power increases greatly to 0.34. This means that the model with all variables included explains about one-third of the variation in the quality of the EPL evaluations. Out of the added variables, only the *type of evaluator* is significant (in line with hypothesis 6). On average, external evaluations score almost two points higher than internal evaluations on the nine-point scale used in this paper. Based on our theoretical framework, the most logical interpretation of this finding is that external evaluators produce EPL evaluations of higher quality because they have more specialized expertise than the Commission's internal evaluators.

Four individual quality aspects correlate significantly with the type of evaluator.[12] These criteria are listed here together with the proportion of external evaluations versus internal evaluations that meets them: (1) a clear operationalization (77% versus 18%), stakeholder consultation (87% versus 73%), triangulation (98% versus 58%) and substantiated conclusions (66% versus 24%). For the other criteria, the difference between both types of evaluators is about 10% or less. Based on these findings, outsourcing evaluations to consultants appears to be particularly useful to produce reports the have high internal validity, a robust methodology and substantiated findings. In other words, hiring external evaluators may contribute to learning by enhancing the quality and the variety of the data that the Commission's EPL evaluations use, as well as the conclusions that are drawn from this information.

The other three technical variables included in the analysis provide no significant explanations. In other words, this paper found no evidence that DGs with more evaluation capacity produce better EPL evaluations than other DGs (hypothesis 5), nor do the data show that legislative complexity negatively affects the quality of EPL evaluations (hypothesis 7).

Table 4: Results of OLS regression

	Model 1: Political factors		**Model 2: Political + technical factors**	
	B *(SE)*	**Sig.**	**B *(SE)***	**Sig.**
Constant	6.53 (0.35)	.00	4.67 (0.42)	.00
Politicization	0.42 (0.27)	.12	0.42 (0.23)	.07
EP involvement	-0.91 (0.35)	.01	-0.99 (0.30)	.00
Council voting	-0.18 (0.43)	.68	-0.10 (0.37)	.78
Clause	-0.58 (0.31)	.06	-0.36 (0.26)	.18
Unit			0.30 (0.22)	.19
Guidelines			0.46 (0.23)	.06
Evaluator			1.91 (0.26)	.00
Complexity			-0.00 (0.00)	.35
N	148		148	
F	3.52 (4, 143)		10.44 (8, 139)	
Significance	0.01		0.00	
Adjusted R^2	0.06		0.34	

6. Conclusion

The aim of this paper was to describe and explain the variance in the quality of the Commission's ex-post evaluations that assess legislative effectiveness. To achieve this goal, a dataset of 153 ex-post legislative (EPL) evaluations was analysed with the help of OLS regression, to test hypotheses derived from two different theoretical views on evaluation quality: a political and a technical perspective.

The descriptive results show that the quality of the Commission's EPL evaluations varies considerably. The average quality score of the reports is 5.6 on a nine-point scale. Most of the evaluations are based on stakeholder input and other types of data, which indicates that their methodology is based on a robust combination of sources. However, the evaluations perform less well regarding the clarity of their scope, the accuracy of their data and the foundations underpinning their conclusions. The worst aspect of the evaluations' quality is their replicability: less than one-third of the reports contained all the material required to repeat their research.

The explanatory analysis shows that the type of evaluator is a significant explanation for the variation in quality. In other words, external evaluators produce considerably better EPL evaluations than the Commission's internal services, especially regarding the clarity of their operationalization, their use of multiple research methods (triangulation) and the extent to which their conclusions are substantiated. These findings suggest that the expertise of specialized consultants is a key asset to enhance the quality of the evaluations. None of the other factors that we studied were found to be significant in the way that we expected.

This paper has two main limitations. Firstly, it does not prove *why* certain factors influence evaluation quality. For example, do external evaluators deliver more quality because they have more expertise (as was suggested by our theoretical framework) or because they are more independent than internal evaluators? A second limitation of the paper is that it focuses on quality indicators that could be efficiently measured using a standardized scorecard. Therefore, criteria related to evaluation processes or the detailed content of reports were omitted. One way to address these limitations would be to conduct in-depth case studies on a number of specific evaluations, so their quality can be fully analysed and the mechanisms that trigger variation in this regard can be explored.

Notes

[1] Some literature about evaluation capacity stresses that this concept is partly political, as it relates to the value that organizations attach to evaluation (e.g. Nielsen et al., 2011: 326-327). However, this article discusses political influences separately and therefore conceptualizes evaluation capacity as the technical means to evaluate only.
[2] Mayne and Schwartz (2005: 305) label this aspect 'sound analysis', but we prefer the name 'robust methodology', as the former seems more related to the criterion of substantiated findings.
[3] This timeframe aspect of evaluation clauses was not derived from theoretical literature, but from discussions with the Commission's Regulatory Scrutiny Board during May 2017.
[4] ec.europa.eu/smart-regulation/evaluation/search/search.do
[5] https://bookshop.europa.eu/en/home/
[6] SWD(2013)228 and SWD(2012)383.
[7] http://ec.europa.eu/atwork/key-documents/index_en.htm
[8] http://eur-lex.europa.eu/homepage.html
[9] Cohen's Kappa was higher than 0.4 for each indicator, indicating a sufficient degree of intercoder reliability (Neuendorf, 2002: 143). Intercoder reliability was not measured for the indicators of a clear intervention logic and substantiated conclusions.
[10] 'evalu*', 'repo*', 'stud*' and 'research'. We also checked the last five articles of each directive or regulation, where evaluation clauses are most commonly found.
[11] The Evaluation of the Measures under Regulation (EC) No 951/97 and the Fitness Check of the Operation and Effects of Information and Consultation Directives in the EU/EEA Countries.
[12] The correlations of the four listed aspects with the type of evaluator are respectively 0.50, 0.17, 0.55 and 0.35.

References

Adcock R and Collier D (2001) Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *The American Political Science Review* 95(3): 529-546.

Böhme K (2002) Much Ado about Evidence: Reflections from Policy Making in the European Union. *Planning Theory & Practice* 3(1): 98-101.

Boswell C (2008) The political functions of expert knowledge: Knowledge and legitimization in European Union immigration policy. *Journal of European public policy* 15(4): 471-488.

Bovens M, 't Hart P and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford handbook of public policy*. Oxford: University Press, pp. 320-335.

Bunea A and Ibenskas R (2017) Unveiling patterns of contestation over better regulation reforms in the European Union. *Public Administration* 95(3): 589-604.

Bussmann W (2010) Evaluation of legislation: skating on thin ice. *Evaluation* 16(3): 279-293.

Cecot C, Hahn R, Renda A and Schrefler R (2008) An evaluation of the quality of impact assessment in the European Union with lessons for the US and the EU. *Regulation & governance* 2(4): 405-422.

Chelimsky E (2008) A Clash of Cultures: improving the "Fit" Between Evaluative Independence and the Political Requirements of a Democratic Society. *American Journal of Evaluation* 29(4): 400-415.

Coglianese C (2012) *Evaluating the performance of regulation and regulatory policy*. Report to the Organization of Economic Cooperation and Development.

Conley-Tyler M (2005) A fundamental choice: Internal or external evaluation? *Evaluation Journal of Australasia* 4(1): 3-11.

Cooksy LJ and Caracelli VJ (2005) Quality, context and use. Issues in achieving the goals of meta-evaluation. *American Journal of Evaluation* 26(1): 31-42.

Cooksy JM and Mark MM (2012) Influences on evaluation quality. *American Journal of Evaluation* 33(1): 79-89.

Datta L (2011) Politics and evaluation: more than methodology. *American Journal of Evaluation*

32(2): 273-294.

Delahais T (2014) *Ex post evaluation of regulation and regulatory policies - The case of EU regulation (ECPR conference paper)*. Available at: http://reggov2014.ibei.org/bcn-14-papers/53-71.pdf (Accessed 24 April 2018).

European Commission (2007) *Responding to strategic needs: Reinforcing the use of evaluation [SEC(2007)213]*. Brussels: European Commission.

European Commission (2010) *Multi-annual overview (2002-2009) of evaluations and impact assessments.* Secretariat-general, May 2010. Available at: http://ec.europa.eu/dgs/secretariat_general/evaluation/docs/multiannual_overview_en.pdf (Accessed 10 July 2015).

European Commission (2013) *Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions. Strengthening the foundations of smart regulation: improving evaluation [COM(2013)686]*. Brussels: European Commission.

European Commission (2015) *Better Regulation Toolbox [SWD(2015)111]*. Brussels: European Commission.

European Commission (2016) *Communication from the Commission to the European Parliament, the European Council and the Council. Better Regulation: Delivering better results for a stronger Union [COM(2016) 615 final]*. Brussels: European Commission.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

Forss K and Carlsson J (1997) The quest for quality - or can evaluation findings be trusted? *Evaluation* 3(4): 481-501.

Häge FM (2007) Committee Decision-Making in the Council of the European Union. *European Union Politics* 8(3): 299-328.

House ER (2008) Blowback: consequences of evaluation for evaluation. *American Journal of Evaluation* 29(4): 416-426.

Kaeding M (2006) Determinants of transposition delay in the European Union. *Journal of Public Policy* 26(3): 229-253.

Lee N and Kirkpatrick C (2004) *A Pilot Study of the Quality of European Commission Extended Impact Assessments*. Impact assessment research centre (working paper).

Majone G (1999) The regulatory state and its legitimacy problems. *West European Politics* 22(1): 1-24.

Mastenbroek E, Van Voorst S and Meuwese A (2016) Closing the regulatory cycle? A meta-evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy* 23(9): 1329-1348.

Mayne J and Schwartz R (2005) Assuring the quality of evaluative information. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction Publishers, pp. 1-17.

Miles J and Shevlin M (2001) *Applying regression and correlation: a guide for students and researchers*. London: Sage.

Neuendorf K (2002) *The content analysis guidebook.* Thousand Oaks, CA: Sage.

Nielsen SB, Lemire S and Skov M (2011) Measuring evaluation capacity: Results and implications of a Danish study. *American Journal of Evaluation* 32(3): 324-344.

Pollack MA (2008) *Member-State Principals, Supranational Agents, and the EU Budgetary Process, 1970-2008*. Paper prepared for presentation at the Conference on Public Finances in the European Union, sponsored by the European Commission Bureau of Economic Policy Advisors, Brussels, 3-4 April 2008.

Poptcheva EM (2013) *Library Briefing. Policy and legislative evaluation in the EU*. Brussels: European Parliament.

Radaelli CM and Meuwese ACM (2010) Hard questions, hard solutions: Proceduralisation through impact assessment in the EU. *West European Politics* 33(1): 136-153.

Rasmussen A and Toshkov D (2010) The Inter-institutional Division of Power and Time Allocation in the European Parliament. *West European Politics* 34(1): 71-96.

Regulatory Scrutiny Board (2018) *Regulatory Scrutiny Board - annual report 2017*. Brussels: European Commission.

Renda A (2006) *Impact assessment in the EU: The state of the art and the art of the state.* Brussels: CEPS.

Rossi PH, Lipsy MW and Freeman HE (2004) *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Sanderson I (2002) Evaluation, policy learning and evidence-based policy making. *Public Administration* 80(1): 1-22.

Scharpf FW (1999) *Governing in Europe: Effective and democratic?* Oxford: University Press.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 67-85.

Steunenberg B (2006) Turning swift policymaking into deadlock and delay: National policy coordination and the transposition of EU directives. *European Union Politics* 7(3): 293-319.

Steunenberg B and Rhinard M (2010) The Transposition of European Law in EU Member States: Between Process and Politics. *European Political Science Review* 2(3): 495-520.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation*. New Brunswick: Transaction, pp. 407-424.

Torriti J (2010) Impact assessment and the liberalization of the EU energy markets: Evidence-based policy-making or policy-based evidence-making? *Journal of Common Market Studies* 48(4): 1065-1081.

Van Golen T and Van Voorst S (2016) Towards a regulatory cycle? The use of evaluative information in Impact Assessments and ex-post evaluations in the European Union. *European Journal of Risk Regulation* 7(2): 388-403.

Van Voorst S (2017) Evaluation capacity in the European Commission. *Evaluation* 23(1): 24-41.

Van Voorst S and Mastenbroek E (2017) Enforcement tool or strategic instrument? The initiation of ex-post legislative evaluations by the European Commission. *European Union Politics* 17(4): 640-657.

Varvasovszky Z and Brugha R (2000) How to do (or not to do) a stakeholder analysis. *Health Policy and Planning* 15(3): 338-345.

Vedung E (1997) *Public policy and program evaluation*. New Brunswick: Transaction.

Versluis E, Van Keulen M and Stephenson P (2011) *Analyzing the European Union Policy Process.* Houndmills: Palgrave MacMillan.

Weiss CH (1993) Where Politics and Evaluation Research Meet. *American Journal of Evaluation* 14(1): 93-106.

Zwaan P, Van Voorst S and Mastenbroek E (2016) Ex-post regulatory evaluations in the European Union: questioning the use of evaluations as instruments for accountability. *International review of administrative sciences* 82(4): 674-693.

Chapter 6: Towards a regulatory cycle? The use of evaluative information in impact assessments and ex-post evaluations in the European Union

Thomas van Golen and Stijn van Voorst

Published as: Van Golen T and Van Voorst S (2016) Towards a regulatory cycle? The use of evaluative information in impact assessments and ex-post evaluations in the European Union. *European Journal of Risk Regulation* 7(2): 388-403.

Abstract

As a part of its Better Regulation agenda, the European Commission increasingly stresses the link between different types of regulatory evaluations. Predictions made by impact assessments (IAs) could be verified during ex-post legislative evaluations, while ex-post evaluations in turn could recommend amendments to be studied in future IAs. This article combines a dataset of 309 ex-post legislative evaluations (2000-2014) and a dataset of 225 IAs of legislative updates (2003-2014) to show how many ex-post evaluations of the Commission use IAs and vice versa. This way, it explores if the Commission's rhetoric of a 'regulatory cycle' holds up in practice. Building on the literature of evaluation use, we formulate the hypotheses that the timeliness, quality and focus of the IAs and evaluations are key explanations for use. Our results show that so far only nine ex-post evaluations have used IAs of EU legislation, while 33 IAs have used ex-post legislative evaluations. Using fuzzy-set Qualitative Comparative Analysis, we find that timeliness is a necessary condition of the use of ex-post evaluations by IAs, suggesting that for the regulatory cycle to function properly, it is crucial to complete an ex-post evaluation before an IA is launched. Future research could repeat our analysis for evaluations of non-regulatory activities or study the causal mechanisms behind our findings.

1. Introduction

During the last sixteen years, the European Commission (EC) has continuously stressed the need to improve its regulatory framework (European Commission, 2000; 2007; 2010; 2013), an ambition which it reconfirmed most recently in its new guidelines for 'Better Regulation' (European Commission, 2015a). By reducing the regulatory burden imposed on citizens and updating the legislation which remains in force, the Commission claims to promote a competitive economy and increase the legitimacy of the EU (European Commission, 2015a: 4).

Impact assessments (IAs) and ex-post legislative evaluations are two key elements of this Commission-wide agenda for Better Regulation[1], as they are tools which can help to make legislation more 'evidence-based' (European Commission, 2015a: 7-9; Radaelli and Meuwese, 2010: 137-140). IAs are studies of the potential costs and benefits of new legislation and other major proposals (European Commission, 2002a: 2), while ex-post legislative evaluations study regulations and directives after they have been in effect for some time (Fitzpatrick, 2012: 478).

In its communications on Better Regulation, the Commission has increasingly stressed the need for a 'regulatory cycle' in which IAs build on the results of ex-post evaluations and vice versa to promote policy learning (European Commission, 2013a: 2-3; 2015b: 288). For example, ex-post evaluations can test if the predictions of IAs have come true, advising the repeal of legislation which has not achieved the predicted effects. In turn, IAs can study the costs and benefits of the amendments which are meant to solve the problems exposed by ex-post evaluations (DG INFSO, 2011: 17; DG MARKT, 2008: 51).

The question, however, is to what extent the rhetoric of a regulatory cycle holds up in practice. Despite the fact that the potential of linking IAs and ex-post evaluation was already recognized by the Commission more than a decade ago (European Commission, 2002b: 9), it appears that the two types of evaluation often remain unconnected in practice. In 2007, an external study of the Commission's IA system showed that only six out of twenty IAs (30%) referred to any kind of interim or ex-post evaluation, a number which includes the use of evaluations on unrelated topics (The Evaluation Partnership, 2007: 86). The Impact Assessment Board (IAB) (2013: 7) – an institution which checks the quality of IAs – stated that in 2013 only one out of six IAs (17%) made use of information from ex-post evaluations.

These numbers suggest that IAs and ex-post evaluations are only loosely connected, although there is a notable lack of more systematic data on the topic from both a descriptive and an explanatory viewpoint (Smismans, 2015: 19). This article seeks to fill this gap by presenting quantitative data on the presence of a regulatory cycle in the EU. By linking a dataset of 309 ex-post legislative evaluations with a dataset of 225 IAs of legislative updates, we are able to describe and explain how many ex-post evaluations refer to IAs and vice versa. More formally, we formulate the research question of our article as follows: how often are IAs and ex-post evaluations of EU law used in subsequent corresponding evaluative instruments and how can variance in this regard be explained? By answering this question, we also hope to provide some recommendations on how the Commission could strengthen its Better Regulation agenda.

Answering our research question does not only have a practical purpose, but also helps to improve academic knowledge. While scholars have published extensively on IAs (e.g. Cecot et al., 2008; De Francesco et al., 2012; Meuwese, 2008; Renda, 2006; Torriti and Löfstedt, 2012) and to a lesser extent on ex-post evaluations in the EU (e.g. Fitzpatrick, 2012; Højlund: 2014b; Stern, 2009), the connection between the two has largely remained ignored (Smismans, 2015: 19), particularly from a quantitative viewpoint. This article helps to bridge the gap between both topics. Hopefully, the numbers we present can be a fruitful basis for future work on both IAs and ex-post evaluations in the EU.

The rest of this article is structured as follows. Section two provides background information about IAs and ex-post evaluations in the EU and how the two types of studies can be linked. In section three we present general theories on evaluation use, from which we derive a number of hypotheses about why IAs and ex-post evaluations may or may not build on each other's results. In section four our data collection, operationalization and tools for analysis are presented. Section five presents both the descriptive and the explanatory results of our research. Section six concludes with recommendations on how the link between IA and ex-post evaluation could be improved and studied further.

2. Impact assessment and ex-post legislative evaluation in the EU

Evaluation in the EU is primarily a decentralized activity in the European Commission. Each Directorate-General (DG) has its own evaluation-related staff and plans its own evaluation reports (Smismans, 2015: 19). While IAs are usually performed internally, ex-post evaluations are often outsourced to external consultants, as the Commission's staff is too small to perform these studies internally and external evaluations are believed to be more objective (Smismans, 2015: 22). Since 2009 a coordinating function for both IAs and ex-post evaluation lies with the Commission's Secretariat-General (SG) (Smismans, 2015: 7).

The evaluation system of the Commission has its origins in the field of programme evaluation, but from the year 2000 onwards it has increasingly been focused on other types of evaluation as well (Fitzpatrick, 2012: 478). After receiving criticism from other EU institutions for a lack of accountability and transparency in its legislative process, the Commission launched a number of reforms in which evaluation played a key part (European Commission, 2000: 6). It became compulsory for new legislation included in the Commission's work programme and other legislation with clear social, economic or environmental impacts to have an underlying IA (Luchetta, 2012: 562). Furthermore, each IA was required to include a section on future monitoring and evaluation (European Commission, 2002b: 7; 2015a: 49). Since 2006 the quality of IAs was checked by the IAB, which was succeeded by the Regulatory Scrutiny Board in 2015 (European Commission, 2015c).

Although systematic ex-post evaluation of EU law was also promised at the beginning of the new century, this topic only received serious attention in Commission documents from 2007 onwards (Fitzpatrick, 2012: 478). During that year the Commission launched a communication stating that ex-post evaluation should be more integrated in the regulatory cycle to provide for Better Regulation (European Commission, 2007: 10). In its recent policy documents on Better Regulation, the Commission increasingly stressed that closer links between IAs and ex-post evaluations are needed to increase the quality of the entire evaluation system (European Commission, 2012a: 3; 2013a: 4, 2015b: 71). The Commission's High-level group for Better Regulation has made similar remarks (High Level Group for Better Regulation,

2012: 12). It remains unclear, however, how much of this rhetoric about a 'regulatory cycle' holds up in practice.

One particular aspect of the Commission's Better Regulation agenda is the 'evaluate first' principle, which states that an IA for a legislative amendment should always be preceded by an ex-post evaluation of the original regulation or directive (European Commission, 2010a: 5). However, in practice the decision on when to start an IA - and by extension the decision to wait for an evaluation or not - lies in the hands of the policy unit or the inter-service group responsible for the IA process (Hartlapp et al., 2013: 430). The IAB could remark on the lack of references to ex-post evaluations when judging an IA, but this issue alone was unlikely to result in a negative opinion, which was only given in case of critical problems (Meuwese and Gomtsyan, 2015: 483, 490-491). It is possible that the link to ex-post evaluations will become more important now that the IAB has been replaced by the Regulatory Scrutiny Board, as this institution also has a formal task in judging the quality of evaluations and could thus play a role in connecting the two evaluative instruments (European Commission, 2015b: 2-4). However, it remains to be seen how this will work out in practice.

Although the Commission's Better Regulation guidelines do state the need of linking IAs and ex-post evaluations, they do not go into much detail about how this can be achieved (European Commission, 2015a: 30). Two DGs have published guidelines for legislative evaluation which provide more information on this issue. DG MARKT states that IAs can inform evaluators which effects were expected at the time the legislation was made (the intervention logic), which could help to formulate research questions (DG MARKT, 2008: 21). In turn, ex-post evaluations can suggest amendments to existing legislation that can be studied in more detail in future IAs (DG MARKT, 2008: 58). DG INFSO states that IAs can also be useful to find stakeholders, as external actors consulted during an IA should ideally be consulted again during an ex-post evaluation to see if and why their views have changed (DG INFSO, 2011: 17). Furthermore, IAs can provide evaluators with background information on the topic and can notify them of potential data sources, indicators for empirical research and methodological pitfalls (DG INFSO, 2011: 12).

Looking beyond the Commission, the European Court of Auditors stated that ex-post evaluations often describe how policies are implemented in practice, which is useful information when drafting an IA (European Court of Auditors, 2010: 40). In a large-scale academic study, Cecot et al. stated that the outcomes of ex-post evaluations can also be used to judge the quality of the assumptions made by IAs (Cecot et al., 2008: 409). Van Gestel and Vranken (2009: 225-228) put this method into practice by using ex-post evaluations in the Netherlands to check the accuracy of the ex-ante assessments which the Dutch Council of State made for new legislation. The results of their research show that even though ex-ante and ex-post evaluations cannot always be compared in practice, systematically checking the outcomes of ex-ante evaluations with ex-post evaluations can be useful to strengthen ex-ante evaluations (Van Gestel and Vranken, 2009: 225-228).

Despite all these potential ways for IAs to use ex-post evaluations and vice versa, existing research on the relation between the two does not provide a very positive picture (The Evaluation Partnership, 2007: 13). To explain this lack of a 'regulatory cycle' despite the Commission's rhetoric on the importance of this issue, the next section presents a number of hypotheses derived from the literature on evaluation use.

3. Theoretical framework

The connection between IAs and ex-post evaluations can be conceptualised as one specific form of evaluation use (De Laat and Williams, 2014: 168). Therefore, this chapter first presents some general explanations for evaluation use and discusses if they can also be applied to the use of IAs by ex-post evaluations and vice versa. In doing this, we focus on situations where both an IA and an ex-post evaluation about legislation actually exist - if this is not the case, this form of use is of course impossible to begin with.

Evaluation use can be defined as the way in which the results from evaluations feed back into an organisation and its policies (De Laat and Williams, 2014: 168). In the literature on evaluation use there has been an extensive focus on the types of evaluation use (Contandriopoulos and Brousselle, 2012; De Laat and Williams, 2014; Forss and Carlsson, 1997;

Højlund, 2014a; Højlund, 2014b; Loud and Mayne, 2014) and explanations for if evaluation results are used in organisations (Loud and Mayne, 2014: 7). The latter area is the focus of this article, as we are looking for the reasons why IAs may or may not use ex-post evaluations and vice versa. Three key explanations from the literature are discussed below: the timeliness of results, the quality of evaluations and the similarity of focus (De Laat and Williams, 2014: 158-162; Smismans, 2015: 15-22).

A first explanation is the timeliness of evaluation results: an evaluation is more likely to be used if it is published before an important decision-making moment and less likely to be used if it is presented right after (De Laat and Williams, 2014: 158-160; Højlund, 2014b: 29). Case studies about the EU's cross compliance legislation and gender research programmes have shown that timeliness is also important in the context of IAs and ex-post evaluation (Smismans, 2015: 19). If an ex-post evaluation is only published while an IA of a proposed amendment is already being drafted, it is less likely the IA will take the ex-post evaluation into account (Bozzini and Hunt, 2015: 64-65; Mergaert and Minto, 2015: 53). In our study, we hope to find out if the conclusion of these studies holds true when analysing a larger number of cases. Therefore, we formulate the following hypothesis:

Hypothesis 1: IAs conducted after an ex-post evaluation about the corresponding legislation is published are more likely to use this ex-post evaluation than IAs conducted while an ex-post evaluation on the same legislation is still being performed.

When it comes to the use of IAs by ex-post evaluations, timeliness problems are almost impossible to occur, as there is a hard requirement for an IA to be published with a legislative proposal and an ex-post evaluation only takes place once the legislation has been in force for a couple of years. Therefore, hypothesis 1 only goes one way.

A second explanation for evaluation use is evaluation quality. If a report is clearly written and sound methodological choices are made, it is more likely something will be done with its results (De Laat and Williams, 2014: 162). Evaluations of poor quality are unlikely to be used because their results cannot be trusted and may undermine the credibility of the user (De

Laat and Williams, 2014: 162). This explanation could also play a role when it comes to IAs and ex-post evaluations, as it is harder to build an IA or ex-post evaluation on earlier research in case the quality of this research is lacking. For example, even though IAs are supposed to formulate a clear intervention logic specifying causes and outcomes (European Commission, 2015a: 48), this does not always happen in practice, making it harder for ex-post evaluations to refer back to them (Luchetta, 2012: 571; Smismans, 2015: 18). Therefore, the following hypotheses are formulated:

Hypothesis 2: IAs are more likely to use ex-post evaluations which are of higher quality.
Hypothesis 3: Ex-post evaluations are more likely to use IAs which are of higher quality.

A third potential explanation for the (lack of) use of IAs by ex-post evaluations and vice versa lies in their difference in focus (Smismans, 2015: 17-23). As the legal and practical requirements for IAs and ex-post evaluations differ to some extent, it can be difficult for them to build on each other's results. For example, IAs tend to be more focused on social and environmental effects and have to take future circumstances into account, while for ex-post evaluations this is not the case. In addition, IAs may be focused on one particular piece of legislation where an ex-post evaluation sometimes considers an entire policy field ('fitness checks') or vice versa (Smismans, 2015: 18) and IAs tend to be more focused on coherence rather than effectiveness (Smismans, 2015: 23). As these differences in focus are expected to have a negative impacts on use (De Laat and Williams, 2014: 162), the following hypotheses are formulated:

Hypothesis 4: The use of ex-post evaluations by IAs is affected negatively by differences in focus between the IA and the ex-post evaluation
Hypothesis 5: The use of IAs by ex-post evaluations is affected negatively by differences in focus between the IA and the ex-post evaluation.

Other authors take a more institutional approach to explaining evaluation use. For example, based on a large-scale study of Swiss ex-post evaluations, Balthasar (2009: 226) shows that

internal evaluations are more likely to be used than external ones. However, since the institutional setting is largely the same for all the evaluations considered in this article (for example, IAs are almost always conducted internally and ex-post evaluations are almost always conducted externally), such issues are not relevant for our purpose. The literature on evaluation use also mentions a broad dissemination of results and stakeholder involvement as key variables for explaining the use of evaluation results: the more actors know about a study, the more likely something is done with its outcomes (De Laat and Williams, 2014: 167; EPEC, 2005: 61). However, since we are only talking about the use of evaluations by other evaluations and not about use by external parties, dissemination is not relevant in the context of our study. Stakeholder involvement was included as an aspect of quality, as will be explained when our operationalization is presented in the next section.

4. Methods and data

4.1 Data collection ex-post evaluations

We answered our research questions with the help of two self-constructed datasets. The first dataset contains 309 ex-post evaluations of regulations and directives conducted or commissioned by the European Commission between 2000 and 2014. Since the Commission does not have one clear format for ex-post evaluations, we included reports with very different kinds of names in our dataset (the most common ones being 'evaluation' 'study', 'review', 'staff working document' and 'implementation report') as long as they have the explicit aim to study EU legislation already in force. Background studies to IAs could be included as well, as long as they fulfil this criterion. To limit the dataset to accessible legislative evaluations, we excluded evaluations focusing entirely on spending activities (even if they do have a legal basis) and five reports only available in French. Also excluded were reports that only present the data of other evaluations and studies which only concern foreign countries[2] or the EU institutions[3], as in this case there is no link to the Better Regulation agenda. In other words, the legislation needs to have an effect on citizens, organizations or member states in the EU. In case multiple

evaluations by the same evaluator about the same legislation existed (e.g. annual Commission reports on a certain regulation), only the most recent one was included. Reports to the Council and the EP only presenting the results of other evaluation reports were excluded.

As the Commission's online database of evaluations is known to be incomplete (Smismans, 2015: 13), the reports were gathered from a number of sources: annual and multiannual overviews of evaluations created by the Commission (2010b), the Commission's search engine for evaluations[4], Commission work programmes, the EU bookshop[5], the annexes to Commission's reports on the financial regulation (2012b; 2013b) and lists of evaluations found on websites of the DGs. The data was checked using an existing dataset of the expertise centre Eureval, by running Google searches for evaluations of all major legislation adopted between 1996 and 2010, by searching for background documents of legislation in Eur-lex and by discussing our data-gathering method with an anonymous SG employee. For a more detailed description of the data collection methods, see chapter 2 of this dissertation.

4.2 Data collection impact assessments

The second dataset used for this article contains all 225 IAs related to legislative updates which were published between the start of the IA system in 2002 and 2014. Unlike with ex-post evaluations, the Commission has a complete database of IAs available online.[6] After importing all IAs from this source (956 in total) we manually excluded the ones which are not about legislation established by the Council or the EP.[7] A first selection was made based on the titles of the IA, after which cases of doubt were checked in detail. Furthermore, IAs about legislation aimed at foreign countries or the EU's internal structure were excluded during this phase, for the same reasons we described in the previous section.

All 495 remaining IAs were coded for whether they relate to updates of previous legislation or to an entirely new regulation or directive. As the EU uses various words for legislative updates, we coded all cases termed as 'amending', 'recast', 'revision', 'repealing', 'simplifying' and 'supplementing' as being updates to previous legislation (225 cases in total). The two categories excluded in this way were IAs of 'new' and 'implementing' legislation (270 cases in total). The former category was excluded from our analysis because new legislation

cannot be expected to build on an ex-post evaluation (Impact Assessment Board, 2013: 7). This is not to say that IAs related to new legislation can never use data from ex-post evaluations - in fact, we encountered four cases where this was so - but in these instances the ex-post evaluation was always related to different legislation, so in this case we cannot speak of a regulatory cycle. The 'implementing' category refers to legislation which codifies an existing principle or agreement in the EU's legislative body. As these principles or agreements were not in legislation before and could not have been evaluated ex-post, we excluded them. IAs that contained both new or implementing legislation as well as legislative updates were included and cases of doubt were checked manually.

As a final step, the dataset of ex-post evaluations and the dataset of IAs were cross-referenced to see for which ex-post evaluations an IA on the same legislation was available at its moment of publication and vice versa. To cross-reference the datasets we first had to link each IA to the correct regulation or directive. This was done by searching for both the number of the IA document and the number of the related Commission proposal in the European Parliament's legislative observatory.[8]

4.3 Operationalization

Evaluation use, the outcome we wish to explain, was operationalized as a simple reference to an IA in the text of an ex-post evaluation and vice versa. The advantage of this method is that it allowed for a large-scale quantitative analysis, although it also means we took even very minor cases of use into account. To search for references to IAs in the ex-post evaluations, the evaluations were manually searched for the keywords 'impact ass*', 'ex ante', 'cost-benefit' and 'cost benefit', with the methodology sections of twenty report being read in detail to check if no keywords were missing. All reports where these search terms yielded results were checked manually to see if the references were indeed about IAs. To search the 225 remaining IAs for references to ex-post evaluations, we used the keywords 'evaluat*', 'ex post', 'interim', 'mid term' and 'retrospective'[9], with the section on procedural issues (which often states the IA's sources) of twenty reports being read in detail to see if no references were missing. Again, each hit was checked manually to see if there was an actual reference to an ex-post evaluation.

After finishing these initial searches, we also checked a random sample of ten IAs (out of 225) for the keywords 'study, 'report' and 'review', as these words are often used in the names of ex-post evaluations. These efforts yielded no additional results and since the amount of work required to search every IA for these three keywords would be disproportionate[10], we did not continue this endeavour. In case an IA did not refer to an ex-post evaluation, but we knew an ex-post evaluation was available from cross-referencing our datasets (see the previous section), we also checked any publicly available background studies to the IA for links to ex-post evaluations. This way, we found three additional references to evaluations. Some of these background studies also contained retrospective elements and were included in the sample of ex-post evaluations.

Concerning timeliness (H1), we looked at the number of months between the publication of the ex-post evaluation and the publication of the IA. According to the Commission's (2009: 8) official IA guidelines, conducting an IA takes about 12 months, so we considered the cases where the number of months was twelve or more to be 'timely' and the cases where the number of months was less than twelve to be 'not timely'. Although it is not impossible for an IA to make use of an ex-post evaluation published when the IA is already under way, it is probably more difficult, as the ex-post evaluation will not be taken into account when sources are collected at the very beginning of the IA process (European Commission, 2015b: 29). Therefore, we believe the threshold of twelve months is justified.

As for evaluation quality (hypothesis 2 and hypothesis 3), it is important to note that quality must be grounded in the subject at hand (Widmer, 2005: 43). Since IAs and ex-post evaluations have different purposes to some extent - even if there are also similarities - we believe judging their quality requires different templates. For example, while ex-post evaluations should specify a clear research question, IAs always have the purpose of comparing the impacts of different policy options, meaning they have less need to explicitly state their purpose. Furthermore, while the reports of ex-post evaluations often contain original research, IAs tend to use empirical data from background studies which are not always transparently conducted or publicly available. Therefore, it is hard to judge IAs on issues like case selection or response rate.

To judge the quality of IAs, we used an adapted version of the scorecard created by Cecot et al. (2008: 418). Since IAs are meant to compare the costs and benefits of policy options, this scorecard is focused on the quantification of alternatives. We slightly adapted the scorecard by replacing the criteria of monetized costs and benefits with two different aspects: the presence of stakeholder consultation and the presence of clear references to sources (through footnotes, a methodology section, or some other way). These changes are in line with the recent Commission standards for IAs, which emphasize stakeholder consultation and methodological soundness (European Commission, 2015b: 49-65). These adaptions to the scorecard also ensure that IAs that study subjects which can be quantified but not monetized are not put at a disadvantage. Three IAs dealing only with matters of fundamental rights and minority rights were coded as missing cases on quality, since for these issues even quantification is probably impossible. For all other types of impacts encountered, at least some quantification seemed possible.

For ex-post evaluations, we used a scorecard with eight criteria to judge their methodological quality: the presence of a clear operationalization, a clear research goal or question, an explanation of selected methods, triangulation, a replicable research design, a clear country selection, a clear case selection and a response rate of >50%. All of these criteria were weighted equally, thus creating an 8-point scale for quality (for more details, see chapter 2 of this dissertation). Table 1 summarizes the scorecards used to judge the IAs and evaluations.

Concerning the focus of the IAs and ex-post evaluations (H4 and H5), we used three indicators. First, we looked at the number of legislative acts studied by each IA and ex-post evaluation. While most reports are about a single regulation or directive, some ex-post evaluations focus on multiple pieces of legislation or even entire policy fields (‘fitness checks’), which makes them potentially harder to compare with IAs. Secondly, we looked at the type of research question. Due to the nature of IAs, they are almost always focused on comparing the costs and benefits of new legislation, but for ex-post evaluations the type of research question may vary. We distinguished between evaluations which look at both costs and benefits, evaluations which only look at either costs or benefits, and ex-post evaluation which study neither costs nor benefits (e.g. pure process evaluations). Thirdly, we looked at the type of

impacts on society which were studied in the IA or evaluation. Working inductively by seeing which types of impacts we found in the actual IAs and ex-post evaluations, we distinguished the following nine categories: (1) economic impacts, (2) environmental aspects, (3) employment impacts, (4) health impacts, (5) safety impacts, (6) customer satisfaction impacts, (7) scientific impacts (e.g. academic output), (8) migration impacts and (9) no impacts on society. The Commission itself uses a simpler typology of three kinds of impacts (economic, environmental, and social), but we found this categorization too limited to map all the variation we observed in practice. Therefore, we distinguished between different kinds of social impacts. The final category ('no impacts on society') was used to cover evaluations only looking at transposition.

Table 1: scorecard for the quality of ex-post evaluations and IAs

Scorecard for the quality of IAs	**Scorecard for the quality of ex-post evaluations**
1. At least some costs are stated. 2. At least some costs are quantified. 3. Provides point estimate or total range of costs. 4. At least some benefits are stated. 5. At least some benefits are quantified. 6. Provides point estimate or total range of benefits. 7. A measure if provided for net benefits or cost effectiveness. 8. At least one alternative is considered. 9. Some costs of the alternative are quantified. 10. Some benefits for the alternative are quantified. 11. A measure if provided for net benefits or cost effectiveness of the alternative. 12. Stakeholder analysis was used. 13. Information on sources is consistently provided.	1. An operationalization is present. 2. A clear research aim or question is stated. 3. The methodology is explained. 4. Methodological tools are provided so that the study could be repeated if necessary. 5. Triangulation of methods is applied. 6. The selection of member states if clearly explained. 7. The selection of cases within member states is clearly explained. 8. The response rate of questionnaires and/or interviews is > 50%.
Maximum score: 13/13	Maximum score: 8/8

Information on each of the three indicators for focus was found by reading the introduction and methodology sections of the evaluations and the 'impact' sections of the IAs. For all three of the indicators, a comparison was then made between IAs and ex-post evaluations belonging to the same pieces of legislation to see if they were similar in focus (yes/no).

Both authors of this article coded half of the IAs for which we needed data on quality and focus. Five cases were coded together beforehand to be sure as few differences as possible would occur between the coders, and any cases of doubt were discussed immediately. The quality scores of the ex-post evaluations were taken from previous research, where intercoder reliability had already been checked and found acceptable (Mastenbroek et al., 2014: 223-225), and the focus scores for the dataset of ex-post evaluations were coded by just one researcher. No cases were found where more than one ex-post evaluation was referred to in an IA or vice versa. Table 2 summarizes the operationalization described in this section.

Table 2: operationalization

Variable	**Operationalisation**
IA use	0 = IA is not referred to in text of the ex-post evaluation 1 = IA is referred to in text of the ex-post evaluation
Evaluation use	0 = Ex-post evaluation is not referred to in text of the IA 1 = Ex-post evaluation is referred to in text of the IA
Timeliness	0 = ex-post evaluation was published 12 months or less before the IA. 1 = ex-post evaluation was published more than 12 months before the IA.
IA quality / evaluation quality	See Table 1.
Object focus	0 = IA and ex-post evaluation have a different focus. 1 = IA and ex-post evaluation have a similar focus.
Problem definition focus	0 = IA and ex-post evaluation have a different focus. 1 = IA and ex-post evaluation have a similar focus.
Impact focus	0 = IA and ex-post evaluation have a different focus. 1 = IA and ex-post evaluation have a similar focus.

It should be noted that a potential drawback of our study is that legislation can still be altered significantly by the Council and the EP after an IA was published by the Commission. This could make it harder for ex-post evaluations to use the results of the IA to test if all its predictions have come true. Ideally a control variable would be added to cover this condition, but

unfortunately it was impossible to map which legislation was amended significantly by the EP within a reasonable timeframe. However, even in case the legislation was amended, the ex-post evaluation could still use the IA as a source for background information or background data. Therefore, this issue should only affect more extensive types of use, such as testing predictions made by the IA or using it as a baseline to measure the exact effects of legislation.

4.4 Method of analysis

As will appear from the results below, the number of ex-post evaluations actively using IAs and vice versa is too low to use regression analysis for our explanatory analysis.[11] For this reason, as well as to gain a better view on which combinations of causal conditions can explain the use of IAs by ex-post evaluations and vice versa, we decided to use fuzzy-set Qualitative comparative analysis (fsQCA) (Ragin, 2008: 9). This method provides information about which (combinations of) explanatory conditions consistently generate necessary or sufficient explanations for the outcome (in the case of this article: evaluation use), based on the proportion of cases which score on both the causal conditions and the outcome.

Since most of our explanatory are coded binary, they can be used in fsQCA without problem. However, evaluation quality and IA quality are ordinal in nature and have to be transformed to vary between zero and one (Ragin, 2008: 85). For such self-constructed scales where little theoretical knowledge exists, this can simply be done by coding the lowest possible score as zero and the highest possible value as one, with the score in-between being the half-way point and the other scores being adapted proportionally (Kogut et al., 2004: 123). As is common practice in fsQCA, we also included the negation of each condition in our analysis to check if any relations work in the opposite way of what we expected (Ragin, 2008: 36).

5. Results

5.1 Ex-post evaluations: descriptive statistics

Out of the 309 ex-post evaluations in our dataset, an IA on the same legislation was available in sixty cases. Out of these sixty evaluations, only ten reports (17%) used the IA which was available on their topic, with a further five reports making use of an IA related to different legislation. This means fourteen cases of using an IA were found in total. Seven of these cases used the IA as a source for background information on their topic, five used data from the IA as evidence to draw conclusions from, three actively tested the predictions that the IA made concerning the costs and benefits of legislation and two used data from the IA as a baseline to measure the amount of change the legislation has caused. These categories are not mutually exclusive. One evaluation did not specify how the IA was used in any way.

Out of the sixty ex-post evaluations where an IA was available, forty cases (67%) are from 2012 or later. This makes sense given the fact that IAs were only introduced from 2003 onwards: older ex-post evaluations are often about legislation enacted before this time. All ten references which we found to IAs were also from 2012 and later, so it is too early to make any claims about developments over time.

Furthermore, eighteen ex-post evaluations referred to IAs conducted by national authorities. As EU directives must be transposed into national legislation, member states often change their laws because of EU requirements and can perform their own IAs accordingly. Fourteen of the evaluations referring to national IAs only referred to a report from the UK, which confirms this country's strong tradition in the field of IAs (The Evaluation Partnership, 2007: 12). One report used an IA from Poland, one report used an IA from Finland, one report used an IAs from Cyprus and Malta and one report used IAs from both the UK and Germany.

We also found ex-post evaluations which referred to IAs prospectively. Seven ex-post evaluations included an IA of a proposed amendment within their report, providing full integration of both types of evaluation. Seventeen evaluations provided suggestions for a future IA in their recommendations, proposing specific changes to the legislation of which the costs and benefits would require further study. Furthermore, four ex-post evaluations used

evidence from a previous IA of a legislative amendment to support their point. In these cases, the Commission had tried to amend the legislation before, but the proposal had been rejected by the Council or the EP. However, the IA which was conducted at the time remained available to feed into future ex-post evaluations.

5.2 Impact assessments: descriptive statistics

Out of the 225 IAs related to updating legislation, an ex-post evaluation on the legislation which was being updated was available in 51 cases. 'Available' means the ex-post evaluation was published at the time the IA was completed: ex-post evaluations published after the IA was finished were not counted here[12]. For 33 out of these 51 cases (65%) we found a reference to the ex-post evaluation in the IA. 21 IAs used the ex-post evaluation as a source of background information to describe their problem, 21 IAs used information from ex-post evaluation as evidence and 20 IAs further investigated amendments suggested by an ex-post evaluation. Other forms of use were not found. Again, it should be noted that these different types of use are not mutually exclusive. Two forms of use could be identified in 17 of the 30 IAs, with 6 IAs containing all the three different kinds of use.

Table 3 provides an overview of the percentage of IAs referring to ex-post evaluations per year. The numbers show that there is an increase in the use of evaluations, although to this moment the 30% in 2011 is the highest number.

As mentioned above, there is a formal requirement for IAs to include a section on monitoring and evaluation (European Commission, 2015a: 49). We found this requirement to be applied well in practice, as only four out of the 225 IAs studied contained no information at all on a future ex-post evaluation. However, the IAs that did provide information varied greatly in their level of detail, ranging from an extensive description of possible indicators to a simple statement about if and when an evaluation should take place.

Table 3: number of legislative IAs referring to ex-post evaluation per year

Year	Number of IAs	%
2003	0	0%
2004	0	0%
2005	0	0%
2006	0	0%
2007	0	0%
2008	7	23.3%
2009	1	3.3%
2010	0	0%
2011	9	30%
2012	2	6.7%
2013	8	26.6%
2014	6	20%
Total	33	100%

5.3 Ex-post evaluations: explanatory analysis

Before presenting the results of the explanatory analysis, it should be emphasized that QCA is an asymmetric method. This means that if a certain condition explains a certain outcome, this does not imply that the absence of the condition also explains the absence of the outcome (Ragin, 2008: 102). The outcomes explained in this section and the next one are respectively the use of IAs by ex-post evaluation and the use of ex-post evaluations by IAs, without making any claims on how non-use of either type of evaluation can be explained.

When analysing data with fsQCA, a useful first step is to test if any of the individual causal conditions are either necessary or sufficient for the outcome (Ragin, 2008: 120). Table 4 provides the explanatory analysis for the use of IAs by ex-post evaluations. As the results show, neither the quality of the IA (H3) nor the comparability of focus between the IA and the ex-post evaluation (H5) are necessary or sufficient conditions for use. In other words, the theoretical explanations which we derived from the literature do not appear to explain whether or not ex-post evaluations build on IAs of corresponding legislation.

A simple look at our data reveals similar results, as no clear patterns emerge. Out of the ten cases where we found an IA was used, six scored eight points or more on IA quality, while

the other four scored five points or less. While we found that the IAs generally studied a much broader range of impacts than the ex-post evaluations (this was so in 36 out of 60 cases, or 60%), in particular when it comes to taking employment and environmental aspects into account, this issue shows no clear relationship with the fact if the IA is used. When it comes to object of study, it was the ex-post evaluations which generally had a broader scope. While 11 out of 60 cases (12%) showed an ex-post evaluation which studied multiple pieces of legislation linked to an IA which was related to just one regulation or directive, the opposite situation never occurred. However, this condition also seems unrelated to whether or not an ex-post evaluation puts its corresponding IAs to use. The hypothesis of Smismans that the different scope of IAs and ex-post evaluations hinders the regulatory cycle is therefore not sustained by our data (Smismans, 2015: 17-22).

Table 4: results of QCA analysis for ex-post evaluations. Proportions > 0.80 indicate a causal factor might be a necessary or sufficient condition; proportions lower than 0.80 indicate a causal factor is unlikely to be a necessary or sufficient condition. The tilde (~) represents the negation of a given variable.

Variable	**Proportion cases cause > use of IA (necessary conditions)**	**Proportion cases use of IA > cause (sufficient conditions)**
IA quality	0.60	0.21
~IA quality	0.40	0.14
Comparison of object focus	0.60	0.13
~Comparison of object focus	0.40	0.36
Comparison of problem definition	0.70	0.27
~Comparison of problem definition	0.30	0.10
Comparison of impact focus	0.50	0.23
~Comparison of impact focus	0.50	0.14

Besides looking at the individual relations, fsQCA can also be used to analyse combinations of conditions. However, for the use of IAs by ex-post evaluations, this too yields no significant results. No single combination of conditions consistently leads to the use of IAs – in fact, consistency scores do not go above 0.28, while QCA generally requires a consistency of at least 0.80 to take a closer look at a combination of conditions (Ragin, 2008: 125).

5.4 Impact assessments: explanatory analysis

Table 5 provides the outcome of the explanatory analysis for individual conditions which might explain the use of ex-post evaluations by IAs. As the results show, timeliness (H1) is a necessary condition for use: if an ex-post evaluation is used, we can be almost sure that it was published at least a year before the IA. Out of the 33 IAs which used an ex-post evaluation, only four cases were untimely (12%), while for the 18 IAs which did not use the available ex-post evaluation, this was seven cases (39%). This result is in line with the finding of Bozzini and Hunt (2015: 64-65) and Mergaert and Minto (2015: 53 that for the regulatory cycle to function, evaluations must be available in time. However, timeliness does not seem to be a sufficient condition: even if an ex-post evaluation is published more than a year before the IA, there are still circumstances in which it is not used in the IA at all.

However, if timeliness occurs *in combination* with a number of other conditions, it also appears to be a sufficient condition for triggering evaluation use. The results of the analysis show one specific combination of conditions which is sufficient: the presence of a timely evaluation, the presence of a high-quality evaluation, the presence of a similar focus of IA and ex-post evaluation in terms of the number of legal acts that are studied and the similarity of the problem definition, and the absence of a similar focus in terms of the impacts which are studied. All eight IAs where this combination of conditions occurs score positively on evaluation use and the combination covers about a quarter of the total amount of cases where an IA uses an ex-post evaluation (coverage: 0.23). In other words: if an evaluation is of high quality, is produced in time and looks at the costs and benefits of the same legislation as the IA, it is very likely it will be put to use in one way or another.

Table 5: results of QCA analysis for IAs. Proportions > 0.80 indicate a causal factor might be a necessary or sufficient condition and have their level of significance provided in parenthesis. The asterisk (*) shows a result is in fact significant. Proportions lower than 0.80 indicate a causal factor is unlikely to be a necessary or sufficient condition. The tilde (~) represents the negation of a variable.

Variable	**Proportion cases cause > use of ex-post evaluation (necessary conditions)**	**Proportion cases use of ex-post evaluation > cause (sufficient conditions)**
Timeliness	0.91 (0.046)*	0.71
~Timeliness	0.09	0.27
Quality ex-post	0.57	0.67
~Quality ex-post	0.43	0.58
Comparison of object focus	0.78	0.69
~Comparison of object focus	0.22	0.47
Comparison of problem definition	0.66	0.75
~Comparison of problem definition	0.34	0.48
Comparison of impact focus	0.91 (0.046)*	0.85 (0.650)
~Comparison of impact focus	0.09	0.55

We have no logical explanation for the fact that a focus on similar impacts between IA and ex-post evaluation was absent in each of the eight cases mentioned above. After all, there is no reason why looking at *different* kinds of impacts would contribute to triggering evaluation use. But at the very least we can conclude that IAs and ex-post evaluations do not have to look at similar types of impacts for the ex-post evaluation to be used. Just like we saw in section 5.3 for the sample of ex-post evaluations, our sample of 51 IAs shows that IAs generally look at a broader range of impacts than the ex-post evaluations. This is the case for 31 out of 51 IAs (61%). In particular, IAs tend to look at environmental and employment impacts more often than ex-post evaluations do. Nevertheless, we found eight cases in which an IA used information from a pure process evaluation, despite the fact that such reports do not factor in

the costs and benefits of legislation for society. This shows it is very well possible for an IA to use an ex-post evaluation which approached its topic from a completely different angle.

6. Conclusion

This article started with the question how often IAs and ex-post evaluations of EU law are used in subsequent corresponding evaluative instruments and how variance in this regard can be explained. The Commission's increasing focus on a 'regulatory cycle' as a part of its Better Regulation agenda raises the question whether or not we can observe a link between IAs and ex-post evaluations of EU law empirically. Combining a dataset of all IAs of legislative updates with a dataset of all ex-post legislative evaluations, we have provided a first quantitative assessment of this question.

Concerning the ex-post legislative evaluations, we found that in sixty out of 309 cases an IA on the same legislation was available, but only ten evaluations actually use the IA in their report. Most of these studies used the IA as a source for background information or evidence to support their conclusions, although a small number of evaluations also tested the assumptions made in the IA. Concerning the use of ex-post evaluations by IAs, we found that for 51 out of 225 IAs a prior ex-post evaluation on the same legislation existed, but only 33 of those IAs actually used the available ex-post evaluation in their report. This means the proportion of IAs making use of an available ex-post evaluation (65%) is much larger than the proportion of ex-post evaluations making use of an available IAs (17%). However, even for IAs there are still 35% of the cases where no use is made of an available ex-post evaluation.

One explanation for this difference could be that an IA of a legislative amendment is usually conducted right after an ex-post evaluation of the previous legislation, which means that the memory of the ex-post evaluation is still fresh. Another potential explanation is that IAs are often conducted internally, making it easier for the Commission to stimulate the use of ex-post evaluations than in the opposite situation. A third possible explanation is that it may be harder for ex-post evaluations to use IAs than the other way around, as IAs are often conducted before legislation is amended by the Council and the EP, meaning that their results may have

little connection with the legislation which actually entered force. This problem was already recognized by the EP in their resolution (2010/2016(INI)) on IAs back in 2011. Although the Inter-institutional Agreement on Better Law-making of 2003 stated that for substantive amendments the Council and EP should conduct their own IA, this principle appears not to have been applied in practice. An article in the recent proposal for a new Inter-institutional Agreement between the EU institutions shows that this problem is once again acknowledged: 'the three institutions aim to ensure that information on the impacts of the act as adopted is available, and can be used as a basis for subsequent evaluation work' (European Commission, 2015d: 6). To find out which of the three mechanisms stated above hinders the use of ex-post evaluations by IAs in reality, more research is needed. In any case, a practical recommendation for the Commission is to require external evaluators to state whether they used the corresponding IA in their analysis and why (not).

As for our explanatory analysis, we found that timeliness is a necessary condition for the use of ex-post evaluations by IAs: an evaluation must be published at least a year before the IA, otherwise it is very unlikely to be used. The quality of the ex-post evaluation and the similarity of its focus between IA and evaluation did not turn out to be significant on their own. However, when an evaluation is timely, is of high-quality *and* looks at the costs and benefits of exactly the same legislation as the IA, we can expect is to be used. Since timeliness is so important, a practical recommendation for the Commission is to actively enforce the 'evaluate first' principle which it has emphasized in the last few years (European Commission, 2010a: 5; 2015b: 17).[13] Since IAs can take a year or more to conduct (European Commission, 2009: 75), it can be tempting to already launch the IA process before an ex-post evaluation is completed, but our research shows this is not a good idea if the Commission takes the idea of a 'regulatory cycle' seriously. Starting the IA process only after an ex-post evaluation is finished significantly increases the chance that the evaluation is used and the opportunity is taken to learn from how the regulation or directive has functioned in the past.

For the use of IAs by ex-post evaluations, our analysis did not reveal any (combinations of) conditions which are sufficient or necessary for evaluation use to occur. This indicates that existing explanations about evaluation use may not be adequate to explain this phenomenon.

One alternative could be to look at more political models of evaluation, suggesting that evaluation results may be used only when they are in line with preferred outcomes (Bovens et al., 2008: 320).

Three other possibilities for future research are worth noting. In the first place, due to the Commission's rhetoric about a 'regulatory cycle', our study has been limited to legislative IAs and evaluations. Therefore, quantitative analysis of the relation between ex-ante and ex-post evaluations of spending activities still seems a fruitful field for further investigation. Secondly, while our research focused mostly on the retrospective use of IAs by ex-post evaluations and vice versa, the prospective side could be worthy of further study. For example, follow-up research could study how often the plans for ex-post evaluations which are stated in IAs are executed in empirical reality. This too is an aspect of the so-called regulatory cycle. As a third option, this type of research could be repeated in the near future when the effects of the new Better Regulation Guidelines are in full force. Especially as the focus on the coupling between ex-ante and ex-post evaluative information is stricter enshrined in the Better Regulation Toolbox (European Commission, 2015b: 28, 254). this could lead to a more visible use of the information the evaluative instruments of the EU generate. When repeating this research it could uncover if the relationship between ex-ante and ex-post instruments has become tighter in practice as it is now on paper.

Notes

[1] From 2010 until 2014 the Better regulation agenda was called 'Smart Regulation'. For the sake of consistency, in this article we only use the name Better Regulation, which was used in official communication before 2010 and is used again since 2015.
[2] To implement this principle, we excluded all legislation initiated by the following DGs and services: DEVCO, ECHO, FPI, ENLARG.
[3] With this we refer to any regulation or directive which is only binding for EU civil servants or for the legal behaviour of the institutions of the EU.
[4] ec.europa.eu/smart-regulation/evaluation/search/search.do (accessed 28-09-2015).
[5] Bookshop.europa.eu (accessed 28-9-2015).
[6] This overview of the Commission's IAs is/was available at ec.europa.eu/smart-regulation/impact/ia_carried_out/cia_2014_en.htm (accessed 28-09-2015).
[7] Such as IAs for Communications, decisions, action plans and recommendations.

[8] The legislative observatory can be found at www.europarl.europa.eu/oeil/search/search.do? (accessed 28-09-2015).
[9] The research was done with Adobe Acrobat Reade, using an advanced search on the folders containing the IAs. Folders were divided per year.
[10] Using the three keywords 'report, 'review' and 'study' could generate 10.000+ hits per year. This could lead to roughly 100+ hits per IA.
[11] Since we have five explanatory conditions, we would need fifty positive cases for regression analysis.
[12] Furthermore, note that some IAs refer to ex-post evaluations of individual policy programmes or action plans. Such evaluations were not counted even if the instrument they study has a legal basis.
[13] It should however be noted that the Commission leaves some discretionary room to ignore the 'evaluate first' principle if it is 'justified by political demands on the Commission' according to p. 256 of the toolbox.

References

Balthasar A (2009) Institutional Design and Utilization of Evaluation: A Contribution to a Theory of Evaluation Influence Based on Swiss Experience. *Evaluation review* 33(3): 226-256.

Bovens M, 't Hart P and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford handbook of public policy*. Oxford: University Press.

Bozzini E and Hunt J (2015) Bringing Evaluation into the Policy Cycle CAP Cross Compliance and the Defining and Re-defining of Objectives and Indicators. *European Journal of Risk Regulation* 6(1): 57-66.

Cecot C, Hahn RW, Renda A and Schrefler L (2008) An evaluation of the quality of impact assessment in the European Union with lessons for the US and the EU. *Regulation and Governance* 2(4): 405-424.

De Francesco F, Radaelli CM and Troeger VE (2012) Implementing regulatory innovations in Europe: the case of impact assessment. *Journal of European Public Policy* 19(4): 491-511.

De Laat B and William K (2014) Evaluation use within the European Commission: lessons for the Commissioner. In: Loud ML and Mayne J (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. London: Sage, pp. 147-174.

DG INFSO (2011) *Evaluating Legislation and Non-Spending Interventions in the Area of Information Society and Media*. Brussels: European Commission.

DG MARKT (2008) *DG MARKT Guide to Evaluating Legislation*. Brussels: European Commission.

EPEC (2005) *Study on the use of evaluation results in the European commission*. Brussels: European Commission.

European Commission (2000) *Focus on results: Strengthening evaluation of Commission activities [SEC(2000)1051]*. Brussels: European Commission.

European Commission (2002a) *Communication from the Commission on Impact Assessment [COM(2002)276]*. Brussels: European Commission.

European Commission (2002b) *Communication for the Commission from the President and*

Mrs. Schreyer: Evaluation standards and good practice [COM(2002]2567]. Brussels: European Commission.

European Commission (2007) *Communication to the Commission from Ms Grybauskaité in agreement with the President: Responding to strategic needs: Reinforcing the use of evaluation [SEC(2007)213]*. Brussels: European Commission.

European Commission (2009) *Impact Assessment Guidelines 2009 [SEC(2009)92]*. Brussels: European Commission.

European Commission (2010a) *Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Smart Regulation in the European Union [COM(2010)543 final]*. Brussels: European Commission.

European Commission (2010b) *Multi-annual overview (2002-2009) of evaluations and impact assessments*. Brussels: European Commission.

European Commission (2012a) *EU regulatory fitness [COM(2012)746]*. Brussels: European Commission.

European Commission (2012b) *The evaluation of the Union's finances based on the results achieved [SWD(2012)383]*. Brussels: European Commission.

European Commission (2013a) *Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions. Strengthening the foundations of smart regulation: improving evaluation [COM(2013)686]*. Brussels: European Commission.

European Commission (2013b) *The evaluation of the Union's finances based on the results achieved [COM(2013)228]*. Brussels: European Commission.

European Commission (2014) *Regulatory Fitness and Performance Programme (REFIT): State of Play and Outlook [COM(2014)368 final]*. Brussels: European Commission.

European Commission (2015a) *Better regulation guidelines* [SWD(2015)111]. Brussels: European Commission.

European Commission (2015b) *Better Regulation Toolbox [SWD(2015)111]*. Brussels: European Commission.

European Commission (2015c) *Decision of the President of the European Commission on the establishment of an independent Regulatory Scrutiny Board [COM(2015)3263]*. Brussels: European Commission.

European Commission (2015d) *Proposal for an interinstitutional agreement on better regulation [COM(2015)216]*. Brussels: European Commission.

European Court of Auditors (2010) *Impact assessments in the EU institutions: Do they support decision-making? [Special report no. 3]*. Luxembourg: European Court of Auditors.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

Hartlapp M, Metz J and Rauh C (2013) Linking Agenda Setting to Coordination Structures: Bureaucratic Politics inside the European Commission. *Journal of European Integration* 35(4): 425-441.

High Level Group for Better Regulation (2012) *Ex-post evaluation*. Brussels: European Commission.

Højlund S (2014a) Evaluation use in the organisational context - changing focus to improve theory. *Evaluation* 20(1): 26-43.

Højlund S (2014b) Evaluation use in evaluation systems - the case of the European Commission. *Evaluation* 20(4): 428-446.

Impact assessment board (2013) *Impact assessment board report for 2013*. Brussels: European Commission.

Interinstitutional agreement on better law-making. European Parliament, Council, Commission on better law-making (PbEU 2003, C 321/1).

Kogut B, McDuffie JP and Ragin CC (2004) Prototypes and strategy: assigning causal credit using fuzzy sets. *European Management Review* 1(2): 114-131.

Loud ML and Mayne J (2014) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. London: Sage.

Luchetta G (2012) Impact Assessment and the Policy Cycle in the EU. *European Journal of Risk Regulation* 3(4): 561-575.

Mastenbroek E, Meuwese ACM and Van Voorst S (2014) Naar een regelgevingcyclus? Evaluatie in de Europese Unie. *Regelmaat* 29(4): 212-228.

Mergaert L and Minto R (2015) Ex Ante and Ex Post Evaluations: Two Sides of the Same Coin? The Case of Gender Mainstreaming in EU Research Policy. *European Journal of Risk Regulation* 6(1): 47-56.

Meuwese ACM (2008) *Impact assessment in EU lawmaking*. Alphen aan den Rijn: Kluwer Law International.

Meuwese ACM and Gomtsyan S (2015) Regulatory scrutiny of subsidiarity and proportionality. *Maastricht Journal of European and Comparative Law* 22(4): 483-505.

Radaelli CM and Meuwese ACM (2010) Hard questions, hard solutions: Proceduralisation through impact assessment in the EU. *West European Politics* 33(1): 136-153.

Ragin CC (2008) *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University Press.

Renda A (2006) *Impact assessment in the EU: The state of the art and the art of the state*. Brussels: Centre for European Policy Studies.

Smismans S (2015) Policy Evaluation in the EU: The Challenges of Linking Ex Ante and Ex Post Appraisal. *European Journal of Risk Regulation* 6(1): 6-26.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 67-85.

The Evaluation Partnership (2007) *Evaluation of the Commission's Impact Assessment System*. Brussels: European Commission.

Torriti J and Löfstedt R (2012) The first five years of the EU Impact Assessment system: a risk economics perspective on gaps between rationale and practice. *Journal of Risk Research* 15(2): 169-186.

Van Gestel RAJ and Vranken JBM (2009) Assessing the accuracy of ex ante evaluation through feedback research: A case study. In: Verschuuren J (ed) *The impact of legislation: A critical analysis of ex ante evaluation*. Leiden, Boston: Martinus Nijhoff, pp. 199-228.

Widmer T (2005) Instruments and procedures for assessing evaluation quality: a Swiss perspective. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction, pp. 41-68.

Chapter 7: The (non-)use of ex-post legislative evaluations by the European Commission

Stijn van Voorst and Pieter Zwaan

Published as: Van Voorst S and Zwaan P (2018) The (non-)use of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy*. Epub ahead of print, 20 March 2018. DOI: 10.1080/13501763.2018.1449235

Abstract

The European Commission has repeatedly emphasized that the results of ex-post legislative (EPL) evaluations should be used to improve the quality of its legislative proposals. This article aims to explain the variation in such instrumental use of EPL evaluations by the Commission. Three high-quality EPL evaluations with varying levels of use were studied in-depth to assess the influence of political factors on evaluation use. The results show that, contrary to expectations, EPL evaluations may be used instrumentally even if their recommendations are opposed by important political actors in the legislative process. This article also shows that a lack of salience of the policy field to which an EPL evaluation belongs in the eyes of the Commission could, in combination with the institution's ambition to reduce its legislative output, be a sufficient condition for the non-use of that evaluation.

1. Introduction

In its official communications, the European Commission (2015: 7; 2016: 2) has repeatedly promoted the idea of evidence-based policy: policy decisions should be based on objective information whenever possible. One important source of such information is ex-post legislative (EPL) evaluations: reports that retrospectively assess the functioning of European legislation (European Commission, 2015: 253). Ideally, EPL evaluations generate knowledge that allows the

Commission to make informed decisions about legislative amendments (European Commission, 2015: 254; Fitzpatrick, 2012: 479). In both the academic literature and this article, such use of evaluations to make informed decisions about policy improvement is labelled instrumental use (Cousins and Leithwood, 1986: 346).[1]

Various academics have discussed to what extent the Commission uses evidence instrumentally in practice. Whereas some research reveals that the Commission often uses scientific evidence to improve legislative proposals (e.g. Rimkuté and Haverland, 2015: 433), other studies have shown that its instrumental use of evaluations is limited due to its politicized environment and because of technical constraints (e.g. Böhling, 2014: 118; Boswell, 2008; Højlund, 2014; Torriti, 2010: 1078). Until now, such empirical research about the Commission's use of evaluations has focused on programme evaluations (e.g. De Laat and Williams, 2014; Højlund, 2014) and ex-ante legislative evaluations (e.g. Radaelli, 2007; Torriti, 2010). Conversely, the Commission's EPL evaluations have rarely been studied (but see Fitzpatrick, 2012; Mastenbroek et al., 2016; Zwaan et al., 2016). Therefore, little is known about what factors affect the Commission's use EPL evaluations.

This omission is unfortunate for two reasons. Firstly, due to its lack of financial and communicative tools, legislation is the Commission's main policy instrument (Lodge, 2008: 282). This makes it important to study if and how the Commission's legislative proposals are influenced by sources of knowledge like EPL evaluations. Secondly, EPL evaluations tend to receive more attention from politicians than evaluations of other policies, since legislation affects the entire public and is usually discussed in parliament (Zwaan et al., 2016: 688). In theory, this makes it likely for the instrumental use of EPL evaluations to be affected by political conditions. For these reasons, this article answers the following research question: *'to what extent and how do political conditions affect the European Commission's instrumental use of EPL evaluations?'*

By answering this question, this article contributes to the ongoing debate about the Commission's nature. Originally, the Commission was perceived as a technocratic institution, aimed at impartial problem-solving (Wille, 2010: 1098). Nowadays, the Commission is perceived as (also) a political actor that pursues its own preferences and acts strategically (Hartlapp et al.

2014: 1; Wille, 2010: 1100). This political perspective on the Commission can be linked to a political view on evaluation use (e.g. Contandriopoulos and Brousselle, 2012: 63-64; Cousins and Leithwood, 1986: 347; Johnson et al., 2009: 379; Weiss, 1993: 95-103). Based on these views, it can be expected that necessary conditions for the Commission's use of EPL evaluations are that they do not contradict the preferences of the policy-makers, veto players or interest groups involved in the European legislative process.

To test these expectations, we conducted an in-depth analysis of three EPL evaluations with varying levels of instrumental use. Extensive document analysis and nineteen in-depth interviews with various actors were used to collect data about the impact of various political conditions on the Commission's use of EPL evaluations. Technical explanations for evaluation use were controlled for, which allowed us to better study the impact of these political conditions.

Contrary to our expectations, our results show that the absence of political opposition to specific evaluation results is not a necessary condition for use. Instead, we found a lack of salience of the evaluated policy field, combined with a commitment to limit legislative output, to be a *sufficient condition for non-use*. If the Commission's political top does not prioritize a policy field, it is unlikely to follow-up on recommendations from EPL evaluations that require new legislative initiatives.

2. Theoretical framework

Political explanations

Since the 1970s, the evaluation literature has increasingly discussed how political conditions, next to technical ones, affect evaluation use (Johnson et al., 2009: 385). This literature generally argues that evaluation use is inherently political, as evaluations allocate praise or blame and may result in policy changes. Actors who feel threatened by evaluations may therefore try to prevent their use or to selectively use those results that fit their agenda (Lederman, 2012: 162; Weiss, 1993: 95-98).

The literature discusses several specific political conditions that affect instrumental evaluation use. A first condition is *policy-makers' preferences* (Lederman, 2012: 162; Weiss, 1993: 97-98), with preferences being defined as actors' beliefs about the feasibility and/or appropriateness of policies (Bunea, 2013). Even when an evaluation recommends certain policy changes, policy-makers may oppose these changes on moral grounds (Weiss, 1993: 97-98) or because they doubt their feasibility. Evaluations are often unable to change such deeply rooted policy beliefs and may therefore remain unused (Weiss, 1993: 97-98).

The literature also shows that the political-institutional context of an evaluation affects its use (Cousins and Leithwood, 1986: 354-355; Shulha and Cousins, 1997: 196). Evaluation results are only one type of input that affects decisions about evaluation use: policy-makers are also likely to consider the position of other actors involved in the decision-making process. Since policies often result from complex negotiations between actors, policy-makers may be unwilling to reopen discussions about them when evaluations recommend to do so (Weiss, 1993: 95), even when they do not object to these recommendations in principle. In particular, we expect evaluation results not to be used instrumentally if they oppose the *preferences of veto players*, as policy-makers must always reckon with the views of actors that can formally block their proposals.

Interest groups are another group of actors whose input may affect evaluation use. Such groups may have no formal veto over policy proposals, but they can put pressure on policy-makers to ignore or implement evaluation results, either directly via lobbying or indirectly via the media. To produce a policy that satisfies a wide range of actors, policy-makers may prioritize such *interest group preferences* over evidence from evaluations (Shulha and Cousins, 1998: 198; Weiss, 1993: 95-98).

A further political condition that may affect use is the interest of politicians and civil servants to *protect their financial resources* (Johnson et al., 2009: 385; Weiss, 1993: 95). Policy evaluations often recommend budgetary reallocations. Policy-makers may ignore such evaluation results if they view them as a threat to their own financial position.

A final relevant political condition is the *media coverage* of an evaluation (Weiss 1993, 95). Media coverage can influence the public opinion about the salience of issues, and issues

that are high on the public agenda are likely to be acted upon. Policy-makers are therefore more likely to pay attention to and be influenced by evaluations when they have been covered by the media (Henry and Mark 2003: 303).

Political explanations and the Commission

Although the Commission is officially a neutral institution (Wille, 2010: 1098), research increasingly shows that it (partly) functions as a political actor in reality (e.g. Hartlapp et al., 2014; Wille, 2010: 1100). Concerning evaluations specifically, the Commission has been shown to ignore results from impact assessments when this was required by negotiations with the European Parliament (EP) and the Council (Torriti, 2010: 1078) and to selectively use evidence from programme evaluations that legitimize its pre-existing views (Boswell, 2008: 472). Based on this, we expect political considerations to also affect the Commission's use of EPL evaluations. Below we specify how and to what extent the political conditions discussed above are relevant to explain the use of EPL evaluations by the Commission.

Policy-makers' preferences are expected to matter in the context of this article. In our study, the Commission is the only decision-maker, as it has the sole right of initiative for most EU legislation and is therefore the only actor to decide about the initial follow-up of EPL evaluations (European Commission, 2015: 297-298). We expect the absence of opposing preferences within the Commission to an evaluation's recommendations to be a first necessary condition for use, since the Commission operates on the basis of a political programme (e.g. Juncker, 2014) and may be unwilling to deviate from this programme when evaluation results contradict it. Since the Commission is not a unitary actor (Hartlapp et al., 2014: 2), we will consider the preferences of its two main parts involved in EPL evaluations: the directorate-general (DG) that manages the evaluation and the Commission's political top that ultimately decides about legislative proposals.

The *preferences of veto players* are also expected to matter for the Commission, as there are two actors that can block its legislative proposals: the Council and the EP. The Commission may consider it useless or needlessly provocative to propose legislation that these institutions oppose, even if an evaluation recommends this (Torriti, 2010: 1078). We therefore expect the

absence of opposition to an evaluation's recommendations from the EP or the Council to be a second necessary condition for use.

Interest group preferences may be especially influential in the context of this article, as the Commission actively consults such groups during most EPL evaluations (European Commission 2015: 280). Existing research shows that interest groups influence many of the Commission's decisions, although their success depends on their resources (Bunea, 2013: 567). Whereas it is common that some interest groups oppose an evaluation's recommendations, we expect that the Commission will not implement recommendations that are opposed by *all* major interest groups involved in a topic. This makes the absence of such opposition a third necessary condition for evaluation use.

The interest to *protect financial resources* is presumably irrelevant for our study due to our focus on evaluations of legislation (i.e. non-spending activities). *Media coverage* is also expected to be unimportant for our study, as media coverage is generally low for EU policies - outside of some sensitive policies not discussed in this article (Princen, 2011: 940). However, this expectation about media coverage will be tested in our empirical analysis.

Technical conditions

Besides political conditions, the literature about evaluation use also discusses several technical explanations. 'Technical' explanations refer to the quality of evaluation products and processes. Existing research shows that these factors influence use because policy-makers only trust evaluation results that they perceive as robust (Cousins and Leithwood, 1986: 358; Johnson et al., 2009: 389; Lederman, 2012: 162). Firstly, since evaluations are a form of applied research, their *methodological quality* matters (De Laat and Williams, 2014: 158-160; Johnson et al., 2009: 379). Secondly, the *credibility of the evaluator* is important: policy-makers put more trust in evaluations published by practitioners with a sufficient reputation (Johnson et al., 2009: 379). Thirdly, an evaluation's *relevance* matters: evaluations are only likely to be used if their content is required by potential users (Johnson et al., 2009: 379).

Fourthly, *stakeholder involvement* is important, as policy-makers can be expected to only trust evaluations that consider the views of actors directly affected by the legislation (De Laat

and William: 2014: 165; European Commission, 2015: 280). Finally, *communication quality* matters: the more an evaluator stays in contact with an intended user during and after an evaluation process (preferably informally), the more likely it is that an evaluation's findings will be relevant for the intended user and will therefore be used (Johnson et al., 2009: 379). Thirdly, the *timeliness* of an evaluation matters, as evaluation results can only be used if they are available before important decision-making moments (De Laat and Williams, 2014: 158). As mentioned, these technical conditions will be controlled for in this study.

3. Methods and data

Case selection

Our study is an in-depth analysis of the Commission's use of three specific EPL evaluations. Three steps were taken to select these cases out of a dataset of 313 cases (updated version of the dataset described in chapter 2 of this dissertation).

Firstly, to select evaluations for which use is likely from a technical perspective, we only considered cases that meet the criteria for a 'good' evaluation product and process described above. Concerning *methodological quality*, we only selected evaluation reports containing a clear operationalization and problem definition (internal validity), a representative country selection (external validity) and data triangulation (reliability). Regarding *credibility*, only evaluations by consultants who conducted at least five other EPL evaluations for the Commission were considered, as this indicates that the Commission trusts their work. Concerning *relevance*, we only selected evaluations that recommend clear legislative amendments. Regarding *stakeholder involvement*, we only selected evaluations presenting stakeholder opinions (for details about these quality criteria, see chapter 5 of this dissertation).

Secondly, only evaluations published between 2008 and 2012 were considered. Evaluations from before 2008 were conducted prior to the introduction of the Commission's evaluation procedures from 2007 (Fitzpatrick, 2012: 478), meaning that any findings about such

cases would be outdated. For evaluations published after 2012, it was too likely that decisions concerning their use had not been made yet.

Thirdly, after intensively scrutinizing the 12 remaining cases, three evaluations were selected. In our first case (the seed and plant propagating material (S&PM) evaluation), the Commission's proposal was congruent with all of the evaluation's recommendations (high level of use), in our second case (the consumer protection cooperation (CPC) evaluation) the Commission's proposal mostly followed the evaluation's recommendations (medium level of use) and in our third selected case (the animal welfare evaluation) the Commission took no new legislative action at all, even though the evaluation recommended this (low level of use). The first two cases therefore allow us to study if the absence of opposition by influential actors is a necessary condition for use, and if so, to trace the mechanisms behind this effect. If no such causal relation is found, the comparison with the third case allows us to find other factors conditioning the Commission's instrumental use of EPL evaluations.

The three selected evaluations offered the advantage that they were all initiated by DG Health and Food Safety (SANTE), so their organizational context was held constant. The Commission also recognized all three cases as full evaluations.[2] Details about the three selected cases are provided in Table 1.

Data collection and analysis

We collected our data via document analysis and semi-structured interviews. Commission proposals for legislative amendments (if available) were studied along with any documents leading up to them. Such documents usually included (1) an action plan based on the evaluation's results, (2) a roadmap for legislative reform, (3) a report on stakeholder consultations, (4) an 'inception impact assessment' about the expected consequences of policy options and (5) a legislative proposal together with the final impact assessment (IA) (European Commission, 2015: 297-306). The document analysis allowed us to identify which of the evaluations' suggestions had been followed up by the Commission and which suggestions it had dropped at what moment.

Table 1: details about selected cases

Case number	Evaluation name	Subject	Publication date	Author	Level of use
1	Evaluation of the Community acquis on the marketing of seed and plant propagating material (S&PM evaluation)	12 directives about seed and plant propagating material (S&PM)	November 2008	Arcadia International, Van Dijk, Civic Consulting, Agra CEAS	High
2	External evaluation of the Consumer Protection Cooperation Regulation (CPC evaluation)	Regulation 2006/2004	December 2012	ICF GHK, Van Dijk, Civic Consulting	Medium
3	Evaluation of the EU Policy on Animal Welfare and Possible Policy Options for the Future (animal welfare evaluation)	23 regulations and directives about animal welfare	December 2010	GHK, ADAS UK	Low

Detailed explanations for these decisions were subsequently gathered via interviews, as we required open questions and follow-up questions to determine the preferences of various actors. To avoid the risk of socially desirable answers we interviewed a broad variety of respondents and guaranteed their anonymity.

In total, we conducted 19 interviews, when possible face-to face (eight cases) and when necessary by phone (nine cases) or e-mail (two cases). For each case we spoke to the

Commission's civil servant who had coordinated the evaluation. Regarding the animal welfare evaluation, we also spoke to the Commission's Secretariat-General (SG), as other interviews showed it had been involved in this case. Additionally, for each case, we interviewed two respondents from different parties in the EP and two external stakeholders that had provided input for each evaluation and that represented significantly different interests (respectively small seed producers and large seed producers, consumer organizations and national consumer authorities (NCAs), and animal welfare NGOs and farmers). We did not interview the Council, as this actor did not finish discussing our second and third case at the time of writing. For each case we also interviewed one of the consultants that conducted the evaluation.

Operationalization

Instrumental evaluation use, our outcome variable, refers to the consideration and implementation of an evaluation's recommendations by its intended user to improve policies (Cousins and Leithwood, 1986: 346). Therefore, we checked during the interviews if the Commission (the intended user) had considered the evaluation's recommendations when deciding about future policies. Furthermore, for each major legislative amendment recommended by the three evaluations, we checked via both document analysis and interviews if any subsequent legislative proposal from the Commission implemented this change. To limit our article's scope, recommendations about legislative implementation or minor clarifications to legislation were ignored.

Concerning the political explanatory conditions, the *Commission's policy preferences* were measured by asking our respondents what amendments to the evaluated legislation the Commission considered necessary before and after the evaluation was conducted. Also, for each of the major recommendations identified, we checked if it was controversial for the parts of the Commission involved in the evaluation's follow-up (the managing DG and the Commission's political top) and if/how this had affected the evaluation's use.

Concerning *veto player preferences* and *interest group preferences*, respondents were asked to what extent each recommendation was in line with the views of the EP, the Council and major interest groups and how the Commission had reckoned with these views in its

decisions. 'Major interest groups' were defined as collectives of interests (like producers and consumers) that were consulted during the evaluation. We checked the views stated by respondents with official documents when possible.

Concerning *media coverage*, respondents were asked if the evaluation was covered by any mainstream media up until the Commission's decision about proposing amendments. Additionally, we analysed if the evaluations were covered by *Politico/European Voice*.[3] Finally, respondents were asked if other factors had influenced the evaluation's use.

Our assessment of the evaluations' technical quality was checked by asking the respondents to judge the internal validity (absence of systematic errors), external validity (generalizability), reliability (absence of random errors) and relevance of the final evaluation report, plus the credibility of the evaluator and the extent to which stakeholders had been involved. The *timeliness* of the evaluation was mapped by asking respondents if the evaluation was available to all relevant actors within the Commission when it decided about legislative amendments (Swanborn, 2007: 323). *Communication quality* was operationalized by asking respondents how often the evaluator had in-depth contact with the Commission during the evaluation process and if informal contact was also possible (Swanborn, 2007: 324).

4. Results

Below we first present the assessment of our technical conditions. We then show how each of our cases 'scored' on the political conditions identified above. After summarizing our results at the end of this section, we proceed with an in-depth analysis.

Technical controls

Almost all respondents who remembered the evaluations in detail confirmed that they observed high standards of validity, reliability, the credibility of the evaluator, relevance and stakeholder involvement. The sole exception was the S&PM evaluation: some respondents believed that this evaluation lacked data about small seed producers (interview 1A, 1D) and/or contained some ambiguous recommendations (interview 1A, 1B). However, these remarks only

concerned some specific elements of the evaluation and other respondents did not support these criticisms (interview 1B, 1E).

Concerning *timeliness*, the interviews confirmed that all the evaluations were available to the Commission before it decided about legislative amendments. Regarding *communication quality*, in all three cases, there was frequent formal and informal contact between the Commission and the evaluator. These results confirm that the use of our three evaluations was not impeded by lacking quality.

Furthermore, none of the respondents believed that the results of the evaluations were changed significantly due to pressure from the Commission. The fact that the respondents were promised anonymity and that many of them moved to new jobs since the evaluations were completed lends some credibility to these claims, although we cannot exclude that the Commission may have subtly influenced the evaluations' findings.

Case description 1: S&PM evaluation

The EU's 12 directives on seed and plant propagating material (S&PM legislation) set the criteria that plant varieties must meet before they may be placed on the European market. The legislation aims to level the playing field for seed producers and to improve agricultural productivity by requiring the registration of plant varieties in national and European catalogues. This in turn requires varieties to meet standards on Distinctness, Uniformity and Stability (DUS-criteria) and, in the case of agricultural crops species, standards on Value for Cultivation and Use (VCU-criteria). Furthermore, the legislation requires national authorities to inspect the quality of individual S&PM lots (Arcadia International et al., 2008: 25-26).

As a part of the Commission's Better Regulation Agenda, the EU's S&PM legislation was evaluated in 2008 to suggest how its effectiveness and efficiency could be improved (Arcadia International et al., 2008: 2). Table 2 lists the evaluation's eight recommendations for major amendments and shows which subsequent Commission documents included plans to implement them.

Table 2: follow-up of the recommendations of the S&PM and the CPC evaluation. Grey cells indicate that proposals to implement the recommendations were included in the document.

	Action plan	Roadmap	Consultation report	Inception IA	Impact assessment	Legislative proposal	Secondary legislation
Recommendations S&PM evaluation							
1 Replace 12 directives with one regulation.	■	■	■	N.A.	■	■	
2 Make the rules for uniformity of varieties more flexible for niche markets.	■	■	■		■	■	
3 Make the VCU rules evolve to adapt to types of agriculture developed for specific uses and to test varieties created by new technologies.	■						■
4 Adapt the requirements for the marketing of seeds to defined categories.					■	■	■
5 Identify links between EU seed law and food legislation.	■	■	■		■	■	■
6 Integrate EU plant health and seed legislation.	■	■	■		■	■	
7 Grant CPVO (Commission institution) the ability to check variety denominations and the right to adopt quality requirements for DUS testing.	■	■	■		■	■	
8 Reinforce provisions to inform seed users.		■	■		■		■
Recommendations CPC evaluation							
1 Enhance the Commission's role in the CPC network (p. 17)	N.A.	■	■	■	■	■	
2 Expand the scope of the regulation's annex (p. 9).			■		■	■	
3 Give additional minimum powers of NCAs (p. 13).		■	■	■	■	■	
4 Clarify NCAs mutual obligations (p. 17).		■	■	■	■	■	■
5 Establish procedural standards for applying NCAs minimum powers (p. 13).		■	■	■			■
6 Establish observatory to assist NCAs (p. 18).		■	■				
7 Clarify the aims of the regulation (p. 17).							

The S&PM evaluation represents a high level of use. Respondents from both the Commission and other organizations confirmed that the Commission took the evaluation's findings seriously when deciding about the future of the S&PM legislation (interview 1A, 1B, 1E). Table 2 also shows that the Commission followed up almost all of the evaluation's recommendations in its legislative proposal (COM(2013)262) and the preceding documents. Only recommendations 3 and 8 were ignored in some of these documents, but they were addressed through delegated acts (European Commission, 2013: 5, 34).

Concerning the *Commission's policy preferences*, the Commission already perceived the need to amend the S&PM legislation before the evaluation took place, as member states had notified it of various problems (like lacking harmonization) with the existing directives (interview 1A, 1B). However, according to respondents representing the Commission and the evaluator, the Commission did not have strong preferences about *how* the legislation should be amended (interview 1A, 1B). The Commission viewed the legislative process as a technical matter, using the evaluation to identify potential policy improvements (interview 1A). One other respondent slightly disagreed with this and stated that the Commission had preferences in line with recommendations 1 and 7 before the evaluation was conducted, but mainly because these solutions had already been suggested by stakeholders (interview 1E).

Concerning *veto player preferences,* the EP viewed the Commission's legislative proposal as too beneficial for large seed companies and rejected it in March 2014 (e.g. resolution A7-0112/2014). On the other hand, the Council generally supported the Commission's views. Some countries (like France) objected to replacing twelve directives on different products with one regulation, but overall there was little controversy among the member states (interview 1A, 1B, 1E).

Concerning *interest group preferences,* large seed producers generally supported the legislative proposal (interview 1D, 1E). However, many NGOs representing small and biological seed producers criticized the proposal for how it handled recommendation 2. Most of these NGOs wanted the DUS-criteria to be abolished altogether for niche market seeds (interview 1A, 1B, 1D, 1E). Whereas the Commission's proposal allowed such seeds to be recognized as 'officially recognized descriptions' to which the DUS-criteria would not apply, the NGOs viewed

the criteria and procedures to apply for this exception as too demanding and opposed the existence of any compulsory registration of niche market seeds on principle (interview 1A, 1D, 1E; IFOAM EU Group, 2013: 6-11).

The interviews and the media analysis showed that *media coverage* was entirely absent for this evaluation.

Case description 2: CPC evaluation

The EU's Consumer Protection Cooperation (CPC) Regulation 2006/2004 aims to enhance the enforcement of certain European consumer protection legislation (as listed in the regulation's annex) by increasing cooperation among national consumer authorities (NCAs). For this purpose, the regulation establishes mechanisms through which NCAs can request each other's assistance, including an IT platform for posting alerts. The regulation also establishes the minimum powers that national authorities must have to be able to assist each other. Furthermore, the regulation creates a European network that coordinates the activities of NCAs (the CPC network) (ICF GHK et al., 2012: 4).

Article 21a of the regulation states that it must be evaluated after five years. Accordingly, an external evaluation of the regulation was completed in 2012, which produced seven recommendations concerning major amendments (ICF GHK et al., 2012: 6-18). Table 2 lists these recommendations and shows which subsequent Commission documents included plans for their implementation.

The CPC evaluation represents a medium level of instrumental use. All respondents believe that the Commission seriously considered the evaluation's results when deciding about possible amendments (interview 2A-2E). The legislative proposal published by the Commission in May 2016 (COM(2016)283) ignored the final two recommendations listed above, but included plans to implement the other five.

Concerning the *Commission's policy preferences*, the interviews showed that recommendations 1 and 3 were longstanding priorities of the Commission, as it viewed more coordinated European action as necessary to protect consumers throughout the internal

market. However, the Commission did not have strong preferences regarding the other evaluation results (interview 2C, 2E).

Concerning *veto player preferences*, the EP supported most of the evaluation's recommendations as being helpful to enhance consumer protection (e.g. resolution A8-0077/2017). However, both the EP and the Council put forward amendments to remove the Commission's right to initiate infringements (recommendation 1), as this proposal was viewed as threatening to national sovereignty. Most member states also opposed the proposed expansion of minimum powers (recommendation 3), as these powers may be difficult to handle for smaller NCAs. Furthermore, many countries opposed the content of some proposed minimum powers (like forcing infringers to compensate consumers) because legally moving these powers to their NCAs would be costly for them (interview 2B, 2C, 2E).

Concerning *interest group preferences*, consumer associations supported all the evaluation's recommendations because they viewed them as beneficial for consumer protection (BEUC, 2016). Business associations only objected to the proposed minimum powers to shut down websites. Our media analysis and the interviews revealed that the evaluation received almost no *media coverage* (interview 2B, 2C, 2D, 2E).

Case description 3: animal welfare evaluation

Legislation is one of the EU's instruments to improve animal welfare. Various European directives protect cattle and experimental animals, for example by banning unfriendly farming methods and regulating space allowances, but most other animal types are not covered by existing EU legislation.

In 2006 the Commission published its first animal welfare strategy. In the context of this strategy an external evaluation of the entire EU animal welfare policy was completed in 2010. Our research only concerns the part of the evaluation about legislation, which recommended to consider expanding the scope of EU animal welfare legislation to protect all animal species (GHK and ADAS UK, 2010: 6).

The Commission (2012: 6) followed up on the evaluation with a second animal welfare strategy, which stated that the possibility of new animal welfare legislation should be

considered in 2014. Respondents confirmed that DG SANTE took the evaluation seriously when drafting this strategy and that it would have been willing to take different decisions if the evaluation's results had recommended this (interview 3A, 3D).

However, when DG SANTE prepared an early draft of a legislative proposal in 2014 it was informed by the SG that the proposal should wait until a new Commission would enter office in November. After this happened, the SG told DG SANTE that the existing animal welfare strategy should be fully implemented before new animal welfare legislation could be considered (even though one aspect of this strategy was considering new legislation) (interview 3A, 3D). Most respondents therefore believe the implementation of the evaluation's recommendation to be blocked by the SG (interview 3A, 3C, 3D, 3E), although the SG states that no such decision was formally taken (interview 3F).

Concerning the *Commission's policy preferences*, the animal welfare unit of DG SANTE always supported further measures to improve animal welfare, including legislation. At the top of DG SANTE and in the SG animal welfare legislation was never considered a priority, but there was little active opposition to the idea either before 2014 (interview 3A, 3D).

Concerning *veto player preferences*, both the interviews and various resolutions (e.g. A7-0216/2012) show that the EP strongly supports stricter animal welfare legislation (interview 3A, 3C, 3D). The Council is more divided about the topic, with countries in North(western) Europe generally supporting new legislation and some countries with much agriculture (e.g. Greece) opposing it.

Concerning *interest group preferences*, farmer associations opposed new animal welfare legislation because it could lead to additional costs. Animal rights groups were also sceptical about the idea of an integrated animal welfare law, as they feared it would include more self-regulation and no stricter welfare standards (interview 3D, 3E). Our interviews and media analysis revealed that there was virtually no *media coverage* of the evaluation.

Summary of the cross case comparison

Table 3 summarizes the three cases and their 'scores' on the explanatory conditions. The CPC case has been split into two groups of recommendations that differ in their level of use; in the other cases, the level of use of the recommendations was relatively similar.

Our theoretical framework predicted that the absence of opposition to an evaluation's recommendations from the Commission, the EP, the Council and major interest groups would be a necessary condition for use. However, as Table 3 shows this is not the case. The S&PM and CPC evaluations are two cases where the Commission implemented respectively all and many recommendations in a legislative proposal, despite significant opposition from respectively the EP plus interest groups and the Council. The results do confirm our expectation that media coverage was absent in all cases.

Table 3: overview of results

Evaluation	**Level of use**	**View of Commission on results**	**View of EP on results**	**View of Council on results**	**Interest groups views**	**Media coverage**
S&PM evaluation	High	No strong opinion	Generally opposed	Generally in favour	Divided	Absent
CPC evaluation: recommendations 1-5	High	Generally in favour	Generally in favour	Generally opposed	In favour	Absent
CPC evaluation: recommendations 6-7	Low	No strong opinion	No strong opinion	No strong opinion	In favour	Absent
Animal welfare evaluation	Low	In favour	Generally in favour	Divided	Generally opposed	Absent

To explain these findings, the next section zooms in on the steps in the follow-up process of each evaluation when specific recommendations were included in or discarded from the Commission's plans. For the first two cases, this allows us to see why the predicted mechanisms were not triggered and if other causes can explain their outcomes instead. Our analysis of the

third evaluation allows us to see if similar mechanisms can play a role in cases with low instrumental use.

5. Analysis

Case analysis 1: S&PM evaluation

As described above, the S&PM evaluation was entirely followed up by the Commission despite opposition from the EP and various NGOs regarding the topic of niche markets. To explain this, we must consider the Commission's contact with these actors throughout the follow-up process. DG SANTE's communication with the EP was mostly handled by its top-level civil servants, while the details about the evaluation's follow-up were decided by its plant health unit. This unit received positive feedback on its plans from the member states via the comitology system, but had no contact with the EP. Accordingly, it was surprised when the proposal was rejected by the EP in 2014 (interview 1A, 1C, 1E). Therefore, the mechanism linking opposition by the EP to non-use that we predicted was not triggered.

The upcoming elections of May 2014 and a critical lobby by NGOs representing small seed producers and biological farmers both contributed to the proposal's rejection by the EP (interview 1A, 1B, 1C, 1E). The Commission's plant health unit had been in contact with these NGOs during the follow-up process of the evaluation and had, as mentioned in the case description, made some concessions to their views. However, in general the unit wished to base its proposal on the evaluation and other evidence, which it felt the NGOs did not provide. The Commission also expected that the NGOs would support the proposal in the end because it would be better for them than no change at all (interview 1A). Whereas some NGOs did indeed take this position, others opposed the amendments on principle (interview 1D). In conclusion, the Commission did not reckon with political opposition from the EP and significant interest groups because it was relatively unaware of the former and it (falsely) thought it could pacify the latter.

The Commission could have relaunched the proposal after its rejection, as some further concessions to the NGOs and the EP may have increased its chances (interview 1A, 1E).

However, this option was complicated by the fact that seed legislation had no direct link to the priorities of the new Juncker Commission (the economy, human rights, migration) (Juncker, 2014). Strict procedural requirements would therefore apply to any new proposal (e.g. a new impact assessment would need to be produced), for which the plant health unit does not currently have the resources (interview 1A, 1C).

Case analysis 2: CPC evaluation

As described above, most of the recommendations of the CPC evaluation were implemented in a legislative proposal from the Commission despite significant opposition from especially the Council. As one respondent stated, the Commission's proposal was 'highly ambitious' because it deliberately ignored objections from the member state (interview 2C). The explanation for this is that the Juncker Commission considered consumer protection a key priority to encourage the European economy (Juncker, 2014: 6). When this Commission entered office it dropped many nearly completed legislative proposals to demonstrate its commitment to its Better Regulation Agenda, but the fledgling CPC proposal continued because it was considered a high priority (interview 2C, 2E).

Conversely, Table 3 also shows that the last two recommendations of the CPC evaluation were ignored by the Commission despite not going against the preferences of any influential actors. How can this be explained? Recommendation 6 (creating an observatory) was still mentioned by the Commission's documents in mid-2014, but had been dropped by October 2015 (during which period the Juncker Commission entered office). This change was solely caused by budgetary reasons: unlike most other recommendations of EPL evaluations, establishing an observatory would cost much manpower to implement. The Juncker Commission had to reduce its civil service from the outset, and any remaining extra capacity for consumer protection was envisaged to be spent on the Commission's increased role in the CPC network (interview 2A, 2B, 2C). This situation appears to be a rare case where the interest to *protect financial resources*, which we predicted to be unimportant in our theoretical framework, does affect the use of EPL evaluations.

Recommendation 7 was not followed up because the Commission considered it to be contradictory: the evaluation first states that the objectives of the regulation must be clarified, but then states that its current objectives are 'appropriate and relevant' (interview 2A). Other respondents also read this recommendation in various ways, confirming its indistinctness (interview 2B, 2C, 2E).

In conclusion, the fact that some recommendations of the CPC evaluation were not followed up by the Commission is best explained by their exceptional characteristics rather than by any fundamental opposition from political actors. Conversely, the recommendations that were relatively controversial have all been followed up because the Juncker Commission viewed them as essential to its political priorities.

Case analysis 3: animal welfare evaluation

As was discussed above, the animal welfare evaluation's recommendation to consider legislative changes was not followed up in the end, despite the fact that it was supported by the responsible Commission DG and the EP. Based on Table 3, an intuitive explanation for this lack of use seems to lie in the opposition of various member states and interest groups.

However, for three reasons, none of the respondents believe that this opposition was influential. Firstly, the idea of new animal welfare legislation was blocked by the SG in 2014, while member states and interest groups only seem to have lobbied about this topic at the DG-level during that time (interview 3D, 3E, 3F). Secondly, various respondents believe that an integrated animal welfare law could have been "sold" to sceptical countries if it had been presented as a simplification effort, with controversial discussions about stricter welfare standards being moved to the comitology system (interview 3A, 3B, 3D). Thirdly, all interest groups state that they were much surprised when the idea of new animal welfare legislation was dropped in late 2014 (interview 3D, 3E).

So what *does* explain the lack of use of the animal welfare evaluation? As in the two other cases, the answer lies in the Juncker Commission's tendency to focus on its political priorities: the economy, human rights and migration (Juncker, 2014; interview 3A, 3B, 3C, 3D, 3E). To demonstrate its commitment to its Better Regulation Agenda, the Commission dropped

many proposals that had no link to these topics, including the draft proposal for new animal welfare legislation (interview 3C, 3D).

In conclusion, the choice not to propose new animal welfare legislation had less to do with political preferences concerning the specific topic and more with general shifts in the Commission's priorities, although according to some respondents the fact that animal welfare was already considered relatively unimportant by the top of DG SANTE and the SG may also have contributed (interview 3A, 3D). The evaluation could not change this situation, as such reports are almost never read at the top of the Commission (interview 3F).

6. Conclusion

This article started with the question to what extent and how political conditions affect the European Commission's instrumental use of EPL evaluations. Based on nineteen in-depth interviews and extensive document analysis, we traced possible reasons for variation in the levels of use of three evaluations that were all of high technical quality.

Our expectation that the absence of opposition to an evaluation's recommendations from major political actors is a necessary condition for their use by the Commission was falsified by our findings. In our first two cases, the Commission implemented all or most of the evaluations' recommendations, despite significant opposition from actors like the EP, interest groups and the Council. In the first case, the Commission was unaware of the EP's opposition and falsely thought it could pacify interest groups with concessions; in the second case, the Commission considered legislative changes too important to reckon with the Council's opposition. In our third case, opposition to the evaluation's findings from interest groups and the Council hardly seemed to have influenced the Commission's decision to ignore its results.

Instead, we found that a lack of salience of the policy field to which an EPL evaluation belongs in the eyes of the current (Juncker) Commission appears to be a sufficient condition for non-use. In other words, if the evaluated legislation has no direct relation to one of the Commission's priorities, the institution is reluctant to propose amendments even when an evaluation recommends this. Our second case fits well with the Commission's economic

priorities and therefore received a legislative proposal, whereas our third case did not and therefore received no follow-up. In our first case, a legislative proposal was already dropped before Juncker entered into office, but attempts to relaunch this proposal were also hindered by the fact that seed legislation is no political priority.

What do these findings imply about the instrumental use of EPL evaluation in the Commission and in general? As our theoretical section explained, the existing literature on instrumental evaluation use (in the Commission and in general) describes various political factors which may impede such use, like the prevalence of pre-existing policy beliefs and the need to safeguard compromises. However, this existing literature pays little attention to the fact that political actors may also have a symbolic interest to reduce their policy output. Since EPL evaluations often recommend changing legislation to improve it, they essentially request policy-makers to frequently propose legislative amendments. In the case of the Commission, such recommendations contradict its plans to propose little legislation outside of its priority fields. This contradiction leads to reduced possibilities for evaluation use.

Our findings suggest that this political interest to limit legislative proposals should be considered when studying the Commission's instrumental use of evidence. As national executives may also commit themselves to limit their legislative output in the context of better regulation agendas, this condition may also be relevant for explaining other policy-makers' use of EPL evaluations.

Our study has two noteworthy limitations. Firstly, due to our focus on the Commission, we did not systematically assess the wider impact of EPL evaluations on legislative outcomes. Our first case showed that even when EPL evaluations affect the Commission's legislative proposals, they may not influence the final outcomes of EU legislative processes, as NGOs and other actors that disagree with evaluation results may still lobby against proposals based on them at the Council or EP. For future research, a more in-depth assessment of such processes would be recommended.

A second limitation lies in our case selection. Since we only studied high-quality evaluations, our conclusions may not apply to evaluations that fail to meet certain technical standards. Furthermore, because we only studied evaluations from DG SANTE the

representativeness of our results could be limited, even though the selected cases covered a wide range of policies. For future research, it is therefore recommended to use a larger number of cases to study whether a lack of salience combined with a commitment to reduce legislative output is indeed a sufficient condition for non-use.

Notes

[1] Other types of use often mentioned in the literature are accountability use, conceptual use and strategic use. Because these types of use may be driven by different factors than instrumental use, they are not considered in this article. For a study about the accountability use of EPL evaluations, see Zwaan et al. (2016).

[2] The Commission classified three cases that met all our criteria as 'studies' instead of EPL evaluations. These cases were dropped to avoid unnecessary discussions about terminology.

[3] The following keywords were entered in the search engine of *Politico* (which also shows the results for articles of *European Voice*) at https://www.politico.eu/: 'seed marketing', 'plant seeds' and 'plant propagating' (S&PM case), 'consumer protection' (CPC case) and 'animal welfare' (animal welfare case). We searched for the entire period of time between the initiation of the evaluation and either the legislative proposal or the decision not to propose legislation.

References

Arcadia International, Van Dijk Management Consultants, Civic Consulting and Agra CEAS (2008) *Evaluation of the Community acquis on the marketing of seed and plant propagating material (S&PM)*. Brussels: European Commission.

BEUC (2016) *Strengthening enforcement*. Available at: http://www.beuc.eu/publications/beuc-x-2016-087_ama_strengthening_enforcement.pdf (Accessed 16 February 2018).

Böhling K (2014) Sidelined Member States: Commission learning from Experts in the Face of Comitology. *Journal of European integration* 36(2): 117-134.

Boswell C (2008) The political functions of expert knowledge: Knowledge and legitimation in the European Union. *Journal of European Public Policy* 15(4): 471-488.

Bunea A (2013) Issues, preferences and ties: determinants of interest groups' preference attainment in the EU environmental policy. *Journal of European Public Policy* 20(4): 552-570.

Contandriopoulos D and Brousselle A (2012) Evaluation models and evaluation use. *Evaluation* 18(1): 61-77.

Cousins J and Leithwood K (1986) Current empirical research on evaluation utilization. *Review of Educational Research* 56(3): 331-364.

De Laat B and William K (2014) Evaluation use within the European Commission: lessons for the Commissioner. In: Loud M and Mayne L (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. London: Sage, pp. 147-174.

European Commission (2012) *European Union Strategy for the Protection and Welfare of Animals 2012-2015 [COM(2012)6]*. Brussels: European Commission.

European Commission (2013) *Proposal for a regulation of the European Parliament and of the Council on the production and making available on the market of plant reproductive material (plant reproductive material law) [COM(2013)262].* Brussels: European Commission.

European Commission (2015) *Better Regulation Toolbox [SWD(2015)111]*. Brussels: European Commission.

European Commission (2016) *Better Regulation: Delivering better results for a stronger Union [COM(2016)615]*. Brussels: European Commission.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

GHK and ADAS UK (2010) *Evaluation of the EU Policy on Animal Welfare and Possible Policy Options for the Future*. Brussels: European Commission.

Hartlapp M, Metz J and Rauh C (2014) *Which policy for Europe? Power and conflict inside the European Commission*. Oxford: University Press.

Henry G and Mark M (2003) Beyond use: Understanding evaluation's influence on attitudes and actions. *The American Journal of Evaluation* 24(3): 293-314.

Højlund S (2014) Evaluation use in evaluation systems - the case of the European Commission. *Evaluation* 20(4): 428-446.

ICF GHK, Van Dijk Management and Civic Consulting (2012) *(External) evaluation of the consumer protection cooperation regulation EC/2006/2004*. Brussels: European Commission.

IFOAM EU Group (2013) *Towards more crop diversity - adapting market rules for future food security, biodiversity and food culture*. Brussels: online publication.

Johnson K, Greenseid L, Toal S, King J, Lawrenz F and Volkov B (2009) Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation* 30(3): 377-410.

Juncker J-C (2014) *A New Start for Europe: My Agenda for Jobs, Growth, Fairness and Democratic Change*. Strasbourg: European Commission.

Lederman S (2012) Exploring the necessary conditions for evaluation use in program change. *American Journal of Evaluation* 33(2): 159-178.

Lodge M (2008) Regulation, the regulatory state and European politics. *West European Politics* 31(1-2): 280-301.

Mastenbroek E, Van Voorst S and Meuwese ACM (2016) Closing the regulatory cycle? A meta-evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy* 23(9): 1329-1348.

Princen S (2011) Agenda-setting strategies in EU policy processes. *Journal of European Public*

Policy 18(7): 927-943.

Radaelli CM (2007) Whither better regulation for the Lisbon agenda. *Journal of European Public Policy* 14(2): 190-207.

Rimkuté D and Haverland M (2015) How does the European Commission use scientific expertise? Results from a survey of scientific members of the Commission's experts committees. *Comparative European politics* 13(4): 430-449.

Shulha L and Cousins B (1997) Evaluation use: Theory, research and practice since 1986. *Evaluation Practice* 18(3): 195-208.

Swanborn P (2007) *Evalueren. Het ontwerpen, begeleiden en evalueren van interventies: een methodische basis voor evaluatie-onderzoek (2nd edition)*. Amsterdam: Boom Onderwijs.

Torriti J (2010) Impact assessment and the liberalization of the EU energy markets: Evidence-based policy-making or policy-based evidence-making? *Journal of Common Market Studies* 48(4): 1065-1081.

Weiss CH (1993) Where Politics and Evaluation Research Meet. *American Journal of Evaluation* 14(1): 93-106.

Wille A (2010) Political-bureaucratic accountability in the EU Commission: Modernising the Executive. *West European Politics* 33(5): 1093-1116.

Zwaan P, Van Voorst S and Mastenbroek E (2016) Ex-post regulatory evaluation in the European Union: Questioning the use of evaluation as instruments for accountability. *International Review of Administrative Sciences* 82(4): 674-693.

Chapter 8: Ex-post legislative evaluation in the European Union: questioning the usage of evaluations as instruments for accountability

Pieter Zwaan, Stijn van Voorst and Ellen Mastenbroek

Published as: Zwaan P, Van Voorst S and Mastenbroek E (2016) Ex-post regulatory evaluations in the European Union: questioning the use of evaluations as instruments for accountability. *International review of administrative sciences* 82(4): 674-693.

Abstract

Evaluations may perform a key role in political systems, as they provide a basis for parliaments to hold their executives accountable. This is equally the case in the European Union. Yet, several factors may work against the use of European Union evaluations for accountability purposes. Members of the European Parliament work under great time pressure and executives may have little incentives to produce high-quality evaluations. This article therefore addresses the question of to what extent and when members of the European Parliament use ex-post legislative evaluations. We present an analysis of 220 evaluations, studying how many were referred to in parliamentary questions. Our main finding is that 16% of the evaluations are followed up through questions. However, the parliamentary questions hardly serve accountability aims. Members of the European Parliament mostly use evaluations for agenda setting purposes. The main variable explaining differences in the use of evaluations is the level of conflict between the European Parliament and Commission during the legislative process.

1. Introduction[1]

In response to concerns about the democratic deficit of the European Union (EU) and the lack of accountability of its institutions, increasing attention has been paid to the accountability of the European Commission. When the Santer Commission resigned in 1999 after allegations of fraud and mismanagement, this increasing attention was followed by action: the succeeding Prodi Commission implemented serious reforms to increase the Commission's accountability (Wille, 2010: 194).

To increase its *political* accountability, the Commission has become more tightly connected to the European Parliament (EP) (Curtin, 2007). Various reforms have provided the EP with a range of instruments to hold the Commission accountable (Wille, 2012: 387), even though the executive powers of the Commission are limited, as it is not in charge of the day-to-day implementation of EU policies. The Commission is, however, responsible for initiating and formulating new policies and for monitoring and enforcing implementation, as well as evaluating EU policies.

The increasing powers of the EP have been followed by stronger demands on the Commission to submit reports about EU policies to the EP (Curtin, 2009: 256-257). Especially promising in this respect are *ex-post* evaluations of EU programmes and legislation. Most legislation adopted nowadays includes a requirement for an evaluation, focusing either on the implementation process or actual impacts (Bussmann, 2010: 280).[2] In theory, these evaluations provide rich information on the fulfilment of policy goals and the responsibilities of actors involved (Corbett et al., 2011: 318-319).

Despite the theoretical potential of ex-post evaluations for political accountability, existing research indicates that the EP does not actively use EU programme evaluations. A study by the Commission demonstrates that 'most evaluations are used only by the officials [administrators] directly involved in the implementation of the interventions that are evaluated' (EPEC, 2005, quoted in Stern, 2009). A recent study of the evaluation of the LIFE programme found that use mainly takes place in the Commission, and not in the EP (Højlund, 2014). Also, Bauer (2006: 723) suggests that the Commission uses evaluations mainly to support its own

decision-making; evaluation can help the Commission 'to improve agenda setting and policy drafting'.

Yet, these findings on the limited use of evaluations by the EP could relate to the fact that existing research on evaluation use in the EU focuses on programme evaluations, which mostly focus on the fate of individual programmes and projects in particular member states (Højlund, 2014: 436). Given the EP's role as a legislator, the situation could be different for evaluations of a more regulatory nature, focusing on the fate of EU legislative policies in the member states. The idea that parliaments are more interested in ex-post legislative (EPL) evaluation would be in line with the finding of Bussmann (2010: 280-282) that parliaments increasingly want to know how the legislation they enact is carried out by the executive. At the same time, it must be noted that Impact assessments (IAs), which can be seen as *ex-ante* legislative evaluations, are not often used by the EP either: Poptcheva (2013: 4-5) found that out of 12.000 EP Committee documents in the 2004-2009 parliamentary term, only one document explicitly referred to a Commission IA. A study by the European Court of Auditors (2010: 21) came to similar conclusions.

The scant existing literature thus suggests that the EP hardly uses ex-post programme evaluations and IAs. This article changes perspective, turning to the question of to what extent and under which conditions the EP uses EPL evaluations to hold the Commission politically accountable. To answer this question we analyse the extent to which Members of the European Parliament (MEPs) ask questions based on EPL evaluations. We do so for a period of roughly three parliamentary terms. This way, we seek not only to add to the literature on EU evaluation use, but also to develop the quantitative knowledge base for understanding accountability in the EU, which is currently rather weakly developed in the literature (Brandsma, 2013b).

2. Accountability in the EU

Bovens (2010: 947-948) distinguishes between authors who view accountability as a virtue of individual actors versus authors who view accountability as a mechanism that structures the relation between several actors. In studies of EU accountability, the second view is most

common and often linked to a political concept of accountability based on the principal-agent (P-A) paradigm. Several EU scholars define accountability as a social relation in which an agent is held to account for his actions to a principal (cf. Curtin, 2007, 2009; Curtin et al.: 2010).

In the EU political system, several P-A relationships exist. Most prominently, the member states act as a collective principal that delegates power to the Commission and the Court of Justice (Pollack, 1997: 203). Increasingly though, the Commission is also seen as an agent of the EP (Curtin, 2007; Proksch and Slapin, 2010). This P-A relationship is central to this article.

Accountability, from a P-A perspective, is an important *ex-post* mechanism for a principal to cope with the risk that the agent deviates from the principal's intentions and interests (Blom-Hanssen, 2005: 631; Curtin, 2007: 525; Pollack, 1997: 108), in addition to administrative procedures and mechanisms that limit the scope of the agent's activity *ex-ante*.

According to Curtin (2009: 257) a proper accountability mechanism requires three steps: first, the principal must have sufficient information about the fulfilment of responsibilities by the agent (Stufflebeam and Shinkfield, 2007: 163). In general, legislatures have an information deficit vis-a-vis the executive, given their lack of expertise and resources (Brandsma, 2013a: 4). Partly, they can compensate for this by acquiring general information about the executive's preferences, allowing them to properly judge information they receive from the executive on specific issues. In the case of the EU, this is more difficult because the executive is not directly linked to the legislature through political ties. Accordingly, information exchange between the legislature and executive takes place in a formalized way, for example through parliamentary questions or EP committee meetings (Proksch and Slapin, 2010).

Over time, the EP has strengthened its general information rights (Brandsma, 2013a: 5). The Commission must submit myriad reports to the legislature, including evaluations (Curtin, 2009: 256-257). The 2010 Inter-Institutional Framework Agreement, which describes the Commission's responsibilities towards the EP,[3] includes a number of principles on the exchange of information.

Second, the agent must be given the chance to explain its actions, for example through a debate. In the EU, this occurs when the EP asks Commissioners to appear before its committees

or to make a statement in the plenary (Corbett et al., 2011: 319-320; Curtin et al., 2010: 258-261). Parliamentary questions provide another forum for the Commission to explain its actions.

Third, the principal must be able to sanction or reward the agent in order to steer its behaviour. Here, it is important to note that the EU political system has a formal separation of powers: the College of Commissioners is selected by the European Council, although its appointment is subject to approval by the EP. This separation of powers means that there is no such thing as a 'vote of no confidence' for individual Commissioners as a direct sanctioning mechanism. While the EP can censure the Commission as a whole, the supermajority requirement to do so makes this very hard (Proksch and Slapin, 2010). The options for the EP to sanction the Commission are therefore limited to holding back budgets or blocking or amending its proposals. Alternatively, the EP can ask the Commission to take action via parliamentary questions or resolutions, which often convey a clear political message (Brandsma, 2012: 79).

3. Ex-post evaluation as a tool for accountability

A potential tool for the EP to hold the Commission accountable are ex-post evaluations. Such evaluations may provide the information necessary for an accountability process to function (Stufflebeam and Shinkfield, 2007: 163). Enabling accountability is generally considered one of the three substantive aims of ex-post evaluation in the public sector, next to learning how to improve existing policy and generating more general knowledge of a policy's intervention logic (Bovens et al., 2008: 322; Patton, 2008; Vedung, 1997: 101).

Accountability is also one of the key aims of evaluation at the EU level according to academic observers (Stame, 2008: 124; Stern, 2009: 71). In the words of Versluis et al. (2011: 207), 'evaluation is fundamental to the EU, given the need for accountability … and performance'. The link between evaluation and accountability also transpires from Commission documents. While the Commission originally presented evaluations as a tool to increase policy effectiveness (European Commission, 2001: 10) it gradually stressed the potential of evaluations to enhance accountability. In the Commission's 2004 Evaluation Guide (European Commission, 2004: 13) accountability was named the primary goal of ex-post evaluation.[4] The EP

acknowledges this accountability function. In its 2001 resolution on the Commission's 'White Paper on European governance' (resolution A5-0399/2001) it argued that the Commission needs to be more transparent and *prove its worth* to the European public. The need for the EP to better scrutinize the executive is also underlined in a 'library briefing' for the EP by Poptcheva (2013), who stresses the need for ex-post evaluations to improve accountability.

While there is increasing attention to accountability, EU ex-post evaluation originally focused on the spending activities of policy programmes, in particular in relation to the EU's Structural Funds (Bachtler and Wren, 2006; Levy, 2001; Stern, 2009). The European Court of Auditors has played an important role in this (mainly financial) system for reporting and evaluation (Curtin, 2007: 553; Versluis et al., 2011: 210). While the role of the Court was originally linked to the EP's power of discharge over the EU budget, it soon 'developed into a fully-fledged audit office', focusing on sound financial management (Laffan, 1999: 254).

Since the late 1990s, a more general interest in ex-post evaluation has grown due to an increasing focus on budgetary stringency, effective programme execution and accountability (European Commission, 2007: Annex 1; cf. Bauer, 2006). Nowadays, legislation is also evaluated, although it should be noted that most legislative evaluation takes the shape of IAs, which are carried out ex-ante (Fitzpatrick, 2012: 478; Summa and Toulemonde, 2002: 411). In sum, the Commission has made great investments in its evaluation system (Fitzpatrick, 2012; Højlund 2014; Stame, 2008; Stern, 2009). Even though EPL evaluations are less common than IAs, earlier research (Mastenbroek et al., 2016) has shown that 33% of the EU's major directives and regulations from the period 2000-2002 have been evaluated ex-post.

4. Theoretical framework

Despite the potential of EPL evaluations for accountability, they often remain unused (Fleisscher and Christie, 2009; Patton, 2008). To understand why this is the case, we take the general literature on 'evaluation use' as a starting point, using the seminal review article by Cousins and Leithwood (1986) as suggested by Højlund (2014),[5] who analysed the use of the evaluation of the EU LIFE programme. This article provides a range of explanations for evaluation use, linking

different types of factors to the specific goals of evaluations. Based on this literature we develop four hypotheses about the use of evaluations for accountability purposes.[6] Cousins and Leithwood (1986) distinguish between two broad sets of explanations for evaluation use. The first set, labelled *evaluation implementation*, focuses on the characteristics of the evaluation and evaluation process. These explanations are fairly 'rationalistic' in nature, as they perceive evaluation as a value-neutral process. This set includes the quality, credibility and relevance of an evaluation to a user, as well as the quality of communication between the evaluator and client (Cousins and Leithwood, 1986). The second set of explanations concerns characteristics of the *decision setting* in which an evaluation is used. This set of factors is more political in nature, which is highly relevant for our purpose of analysing political accountability.

Evaluation implementation: rationalistic factors

Our first two hypotheses are grounded in the assumption that evaluation use depends on the characteristics of the evaluation or evaluation process (Cousins and Leithwood, 1986; Johnson et al., 2009). Evaluation use is, first of all, argued to be a function of evaluation quality (Cousins and Leithwood, 1986: 347). Low-quality evaluations make the user vulnerable to criticism, which reduces the chances of use (Cooksy and Caracelli, 2005: 31; EPEC, 2005: 39-40). Meta-evaluations show that the methodological quality of evaluations is not always guaranteed; in some examples less than half of the evaluations meet minimal standards of quality (Datta, 2006: 434; Forss and Carlsson, 1997: 490). The quality of EU evaluations is also disputed (Mastenbroek et al., 2016; Versluis et al., 2011: 224).

While evaluation quality can be an impediment to use by administrators closely involved in the evaluated policy, we believe this to be less of a concern for principals in the case of political accountability; we expect those principals to be interested primarily in the conclusion of an evaluation. To the extent that MEPs care about the quality of the evaluation process, we expect them to mainly care about the objectivity of the results (Poptcheva, 2013). This is in line with Cousins and Leithwood (1986: 347), who argue that the objectivity of an evaluation affects the use of evaluations for accountability purposes. This view is also supported by various other

authors (Rossi et al., 2004: 36; Stufflebeam and Shinkfield, 2007; Vedung, 1997), who argue that evaluations intended for accountability purposes should be conducted by external researchers, who are likely to be more independent. Studies on EU IAs also support this view, suggesting that MEPs consider the independence of evaluations before deciding to use them - often affecting use in a negative way (Poptcheva, 2013). According to Stern (2009: 70-72), there is a widespread belief that the Commission advocates objectivity in its evaluations on paper, but not in practice.[7] Not surprisingly, one of the alleged reasons for the very little use of IAs by MEPs is their distrust in the objectivity of the information (Poptcheva, 2013). On this basis we formulate the following hypothesis for the use of EPL evaluations:

Hypothesis 1: Evaluations of European legislation conducted by external companies are more likely to be used for accountability purposes by MEPs than evaluations that are conducted internally by the Commission.

Furthermore, Cousins and Leithwood (1986: 347) argue that evaluation use is a matter of relevance to the user; evaluations that do not match the goals of their audience are less likely to be used (Toulemonde, 2006). When it comes to using evaluations for accountability purposes, various scholars suggest that evaluations must be aimed at finding out if a policy has worked, rather than how it can be improved: accountability evaluations need to focus on outcomes rather than on processes (Lehtonen, 2005: 170-171; Rossi et al., 2004; 36; Stufflebeam and Shinkfield, 2007: 161). When it comes to holding the Commission accountable, however, this distinction seems less relevant, as the Commission has important responsibilities in particular in supervising the implementation *process*. We therefore expect MEPs to care equally about both outcome and process evaluations.

We do expect MEPs to be affected by the clarity of the evaluation results. Cousins and Leithwood (1986: 347) argue that evaluation use is contingent on *communication quality*: the clarity of the results to the evaluation audiences. In reality, evaluations compete with other sources of information (Weiss, 1993: 94): MEPs are often overloaded with information and have

a hard time deciding what information is important (Linter and Vaccari, 2005: 23). We therefore expect that evaluations, in order to be used by MEPs, must either be short or have concise executive summaries (Forss and Carlsson, 1997: 495). This leads to the following hypothesis:[8]

Hypothesis 2: Evaluations that are short or have a concise executive summary are more likely to be used by MEPs for accountability purposes than lengthy evaluations or evaluations without a concise executive summary.

The decision setting: political factors

The 'evaluation implementation' perspective is fairly 'rationalistic' in nature, working from the conception of evaluation as providing value-neutral information (Bovens et al., 2008: 325). However, this view is disputed in the literature on evaluation use. Evaluation, crucially, is argued to be a political venture: it is 'nothing but the continuation of politics by other means' (Bovens et al., 2008: 321). Arguably, this is especially true in the EU, 'where a plethora of actors and institutions are fighting to champion their own agenda and their preferred course of action' (Versluis et al., 2011: 223). We therefore need to bring in more political explanations - which are found in the realm of the 'decision setting' explanations (Cousins and Leithwood, 1986).

One important characteristic of the decision setting is the importance attached to an evaluation (Johnson et al., 2009: 385). Especially in accountability relationships, this importance is affected by the risks for the principal that the agent does not deliver the task for which it was delegated power ('moral hazard'). Benjamin (2008: 335-336) suggests that the *significance* and the *nature* of this risk provides different incentives for evaluation use.

The accountability relationship is affected, first, by the *significance* of the principal's risk. Here, the basic question is: 'How costly is possible shirking by an agent to the principal?'. According to Benjamin (2008), the significance of the risk increases the need to hold an agent to account, which, in turn, increases the chances of evaluation use. We expect the significance of this risk for MEPs to depend partly on *issue salience*, that is, the relative importance attached to a particular policy by MEPs. As elected politicians have limited time and capacity, they must be

selective and give priority to certain issues. We expect that the salience of a policy issue to MEPs will affect the perceived significance of the risk that this policy does not work, and thus the need to hold the Commission accountable. This leads to the following hypothesis:

Hypothesis 3: The higher the salience of a piece of legislation to MEPs, the higher the chances that an evaluation of the legislation is used by MEPs for accountability purposes.

The importance attached to an evaluation is also affected by the *nature* of a principal's risk. In this case the question is: 'Is there a possible conflict of interest between principal and agent?' (Benjamin, 2008).[9] Here, rivalries and power struggles play a role (Johnson et al., 2009: 385). In this regard, it is important to return to the fact that the EU has a formal separation of powers. This results in a situation where the EP and Commission can act relatively independently of each other to pursue their interests during the legislative process (e.g. Nugent, 2010: 206). In a situation of a conflict of interests, we expect MEPs to be more concerned with holding the Commission accountable. This leads to the following hypothesis:

Hypothesis 4: The more conflict during the legislative process, the higher the chances that an evaluation of the legislation is used by MEPs for accountability purposes.

5. Methods and data

The preceding hypotheses are tested using a dataset of 220 EPL evaluations commissioned or carried out by the Commission (updated version of the dataset used in chapter 2 of this dissertation).[10] The evaluations were gathered from multiple sources: Directorat-General (DG) websites, Commission evaluation documents (European Commission, 2010a), Commission work programmes (European Commission, 2010b), the Commission's evaluation search engine,[11] systematic Google searches and searches for Commission reports on legislation through the simple search procedure in Eur-lex.[12]

Dependent variable

We operationalized our dependent variable 'use of an evaluation for accountability purposes' by turning to parliamentary questions asked by MEPs. The reason for this is that European parliamentary questions (EPQs) have been recognised as one of most visible and easy instruments available to MEPs to hold the Commission accountable (Proksch and Slapin, 2010: 60; Wille, 2010: 60). While the literature suggests that the extent to which parliamentary questions are used for accountability purposes in *national* politics is low because MPs mainly use them for attention, Proksch and Slapin (2010: 60) argue that this is different for the EU. According to them, the second-order nature of EP elections and little media attention for EU politics provide few incentives for MEPs to use questions for personal publicity or to gain votes. Accountability and information are therefore a much more important function of EPQs.

For all EPL evaluations in our dataset, we first searched for corresponding EPQs, using the EP's website on parliamentary questions and declarations.[13] For three parliamentary terms (1999-2004, 2004-2009 and 2009-2014) we searched for EPQs referring to the evaluations in our dataset.[14] The EPQs that resulted from this search were read to make sure that reference was indeed made to the evaluations included in our dataset. For our initial analysis we adopted a broad understanding of accountability purposes. We included questions in which evaluations were used to ask additional information, to demonstrate shortcomings and to give the Commission a chance to explain its actions and responsibility, as well as questions that 'steer' the Commission by summoning it to change existing legislation. In the latter case there is a close empirical link between retrospective and prospective parliamentary scrutiny (Wille, 2010). This broad definition is in line with the existing literature on the quantitative measurement of accountability (Brandsma, 2013b). The dependent variable was measured as a dichotomous variable: an evaluation either has or has not been referred to in EPQs.

Independent variables

Starting with the rationalistic variables, we analysed whether the evaluations were carried out internally by the Commission or carried out by *external researchers* by consulting the title pages

of the evaluation reports gathered. This resulted in a dichotomous variable. To measure the *clarity of the reports*, we assessed whether an executive summary of no more than 10 pages was included in or attached to the report; reports with under 10 pages of main text scored automatically on this condition (Mastenbroek et al., 2016).

Continuing with the political variables, we measured *issue salience* to MEPs by counting the total number of so-called recitals attached to the legislation evaluated. Proposals for EU law always include texts that set out various arguments - recitals - for why regulatory action is needed. It is suggested in the literature that more recitals indicate a higher salience of a piece of legislation (e.g. Warntjen, 2012: 171). This is a continuous variable. The *degree of conflict* was measured by counting the number of amendments proposed by the EP to the original Commission proposal underpinning a piece of law. While not all amendments have (equal) political significance, the total number of amendments is seen as indicative of the conflict between the Commission and the EP (Franchino, 2000: 75). This variable is also continuous. The different variables and their expected effect on the referral to evaluations in EPQs are summarized in Table 1.

Method of analysis

Since our dependent variable is dichotomous (EPL evaluations being referred to or not referred to in EPQs), and our independent variables are either categorical or continuous, we chose to use a binary logistic regression analysis (Long and Freese, 2006) to test our hypotheses.

Table 1: Overview of independent variables

Variable	Indicator	Measurement	Expected effect on use of evaluations
Rationalistic view			
External evaluation	Not relevant	0 = internal evaluation 1 = external evaluation	+
Concise executive summary	Number of pages of executive summary or of report of max. 10	0 = evaluation is longer than 10 pages and does not have a summary of less than 10 pages 1 = evaluation is shorter than 10 pages or contains a summary of less than 10 pages	+
Political view			
Significance of risk	Issue salience	Total number of recitals	+
Nature of risk	Legislative conflict	Number of amendments	+

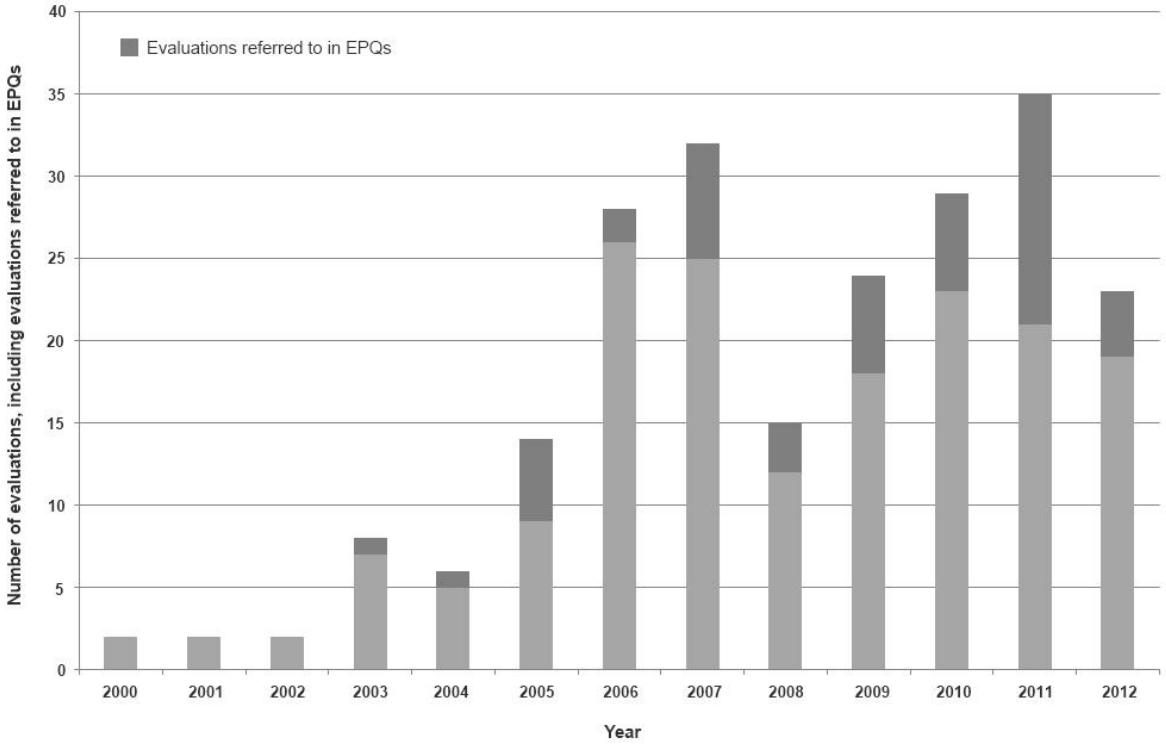
6. Results

Descriptive results

To show to what extent evaluations are used, we first provide some descriptive results. Of the 220 EPL evaluations included in our dataset, 49 (22.3%) were referred to in EPQs. More specifically, 29 evaluations were referred to once, seven were referred to twice, seven were referred to three times and two were referred to four times. Four evaluations appeared even more often: Immigration and Asylum legislation (five times); the Tobacco Products Directive (eight times); animal welfare legislation (10 times); and, finally, the Data Retention Directive (19 times). In total, 114 questions referring to EPL evaluations were identified.

Figure 1 shows the number of EPL evaluations and evaluations being referred to in EPQs per year. While we see an increase in the number of EPL evaluations being published, there appears to be no clear trend in referral to evaluations in EPQs.

Figure 1: Number of evaluations and evaluation being referred to in EPQs



Explanatory analysis

Turning to the explanatory analysis, we estimated a model containing only the rationalistic variables and a full model also including the political variables However, the loglikelihood ratio tests indicated that both the models were not significant.

Finding these insignificant results for our model, we decided to take a closer look into the operationalization of our *dependent* variable. In line with quantitative accountability studies, we initially included all EPQs referring to an evaluation. To verify whether this is indicative of 'evaluation use for accountability purposes', we looked closer into the types of

questions asked. Working inductively, we identified four types of questions. The first type includes questions about *evaluation follow-up*: 'What has the Commission done or is it doing based on the evaluation report?' (62 questions). The second type of questions concerns *evaluation content*: 'Why are certain topics included, excluded or treated in a specific way in the evaluation?' (19 questions). Two other types of questions are related to the evaluation process, dealing with *evaluation timing*: 'Is the Commission planning an evaluation or when will this evaluation take place?' (27 questions) - and *evaluation stakeholders* - 'Why were certain actors included or excluded during the evaluation process?' (six questions). Remarkably, no questions at all were of a truly retrospective nature, focusing on the performance of past policies or the fulfilment of specific responsibilities by the Commission.

On the basis of the different types of questions, we decided to narrow down the EPQs indicative of 'evaluation use for accountability purposes'. To test our hypotheses, we included only those EPQs that contained follow-up questions about evaluations, believing that these relate the most to the responsibility and behaviour of the Commission in relation to the legislation evaluated. Table 2 presents the results of the new analysis.

The loglikelihood ratio test indicates once again that the model containing only the rationalistic variables is not significant. However, this changes when we add the political factors: in that case, the model becomes significant at the 0.05 level. In the full model, the effect of the presence of a clear executive summary and the number of recitals, while pointing in the expected direction, remains insignificant. This is also the case for the effect of the type of evaluator, which, moreover, points in the opposite direction compared to what we expected. However, we do find that the number of amendments has a significant effect on the use of evaluations: the analysis shows that in terms of predicated probabilities, the chances of an evaluation being used increases by 2.1% for every extra amendment proposed, holding all other variables at their average values. This finding is in line with our hypothesis that evaluations of politically sensitive policies - leading to more conflict during the legislative stage - are more likely to be referred to by MEPs for accountability purposes than less sensitive policies. However, our other political variable, the *significance* of the risk, measured by the number of

recitals, did not have an effect on evaluation use. A possible explanation for this could be that the salience of a policy is less stable than the level of conflict; the number of recitals in a piece of legislation may therefore be less indicative of a policy's significance at the time that an evaluation is published. Another explanation, of course, could be that salience alone is simply not a good indicator of the significance that MEPs attach to the risk that a policy does not work. The analysis thus shows that it is only the level of conflict between the EP and the Commission during the legislative stage that significantly increases the chances that MEPs use ex-post evaluations for accountability purposes.

Table 2: Results of the logistic regression

	Rationalistic model			**Full model**		
	B *(SE)*	**Sig.**	**Exp*(B)***	**B *(SE)***	**Sig.**	**Exp*(B)***
Constant	-1.69 (0.36)	0.00	0.18	-2.18 (0.49)	0.00	0.113
External evaluator	-0.13 (0.39)	0.74	0.88	-0.13 (0.40)	0.74	0.877
Concise summary	0.15 (0.40)	0.71	1.16	0.27 (0.42)	0.53	1.306
Recitals				-0.01 (0.01)	0.34	0.988
Amendments				0.02 (0.01)	0.00	1.016
	N = 205 Chi Square = 0.24 Sig = 0.89			N = 205 Chi Square = 11.41 Sig = 0.02		

When we further analyse the residuals - identifying those cases with a studentized residual greater than 2^{15} - it becomes clear that most of the evaluations that are referred to in EPQs but cannot be explained by our model concern post-material issues such as the protection of consumers and the environment. Table 3 presents these evaluations. This analysis sustains the political view arising from our statistical results.

Table 3: Studentized residuals

	Studentized residual	Amendments	Political Group
Evaluation of the Transport of Dangerous Goods	2.10	12	Greens (1)
Report on Noise Operation Restrictions at EU Airports	2.15	20	ALDE (1), S&D (1)
Evaluation of the Environmental Noise Directive	2.00	36	S&D (3)
Fitness Check Water Policy	2.09	99	ALDE (2)
Study of Directive 2001/29/EC on the harmonization of certain aspects of copyright and related rights in the information society	2.17	58	S&D (1), EPP (1)
Directive 2003/86/EC on the right to family reunification. Synthesis report.	2.20	16	Greens (1) S&D (1)
Report on the application of Directive 94/80/EC on the right to vote and to stand as a candidate in municipal elections by citizens of the Union residing in a Member State of which they are not nationals	2.09	25	ALDE (1)
First Annual Report on Immigration and Asylum	2.00	38	PPE (2), ALDE (1), GUE/NGL (1)
Expert study on the issues arising from a reduced time frame and the options allowed for submitting recapitulative statements	2.40	0	ALDE (2)
Interim Evaluation of the European chemical market after the introduction of REACH	-2.15	351	None
Operation and effects of information and consultation directives in the EU/EEA countries. Fitness check	2.02	35	S&D (1)
Second report on the implementation of regulation on civil aviation security	2.08	14	EPP (1)
Report on the operation of Regulation (EC) no 1185/2003 on the removal of fins of sharks on board vessels	2.04	16	EPP (1)

While it is recognised in the evaluation literature that evaluations are often used for political purposes (e.g. Weiss et al., 2005), various broader models on agenda-setting and policy change are worth turning to for explaining evaluation use. Kingdon's (1995) multiple streams (MS) model, for example, directs our attention to the fact that evaluations in and by themselves do not result in agenda-setting. More commonly, evaluation results need to be picked up by actors in the political stream, for instance, by MEPs asking EPQs, to get the issue onto the EU political agenda (for an application to the EU setting, see, e.g., Ackrill and Kay, 2011). Somewhat differently, the punctuated equilibrium model (Baumgartner et al., 2014) directs our attention to the fact that evaluations can change the policy image around which a so-called policy subsystem is structured. When actors become aware that the current way of looking at a policy problem or solution is flawed, this could result in a destabilization of the policy subsystem (for an application to the EU setting, see, e.g., Princen, 2013). Based on these models, MEPs are expected to ask questions about evaluations when alternative policies are available or when they want to make sure that evaluation results are not ignored.

More in general, the agenda-setting literature could be used to shed light on the questions of why and when EPQs are used for agenda-setting as compared to other pathways of agenda-setting, or under which conditions they are used to specify or expand an issue based on an evaluation outcome (e.g. Cobb et al., 1976; for an application to the EU setting, see Princen and Rhinard, 2006).

7. Conclusion

Although various EU institutions stress the potential of ex-post evaluations for holding the Commission accountable, little is known about the actual use of ex-post legislative (EPL) evaluations for accountability purposes. By analysing EPQs that explicitly refer to EPL evaluations, we developed a measure of accountability activities, and the use of ex-post evaluations in this process. Our analysis showed that out of 220 evaluations analysed, 49 (22%) evaluations were referred to in EPQs, of which 34 (16%) evaluations were used to steer the behaviour of the Commission. This percentage is above expectations when we compare it to the

use of EU IAs by MEPs and when we consider that research on evaluation use is difficult as MEPs may not always explicitly refer to evaluations in their questions.

At the same time, our analysis revealed a clear forward-looking agenda setting outlook rather than a backward-looking attitude of MEPs when it comes to making use of EPL evaluations. Surprisingly, we found no retrospective use of evaluations at all: instead, MEPs go beyond demonstrating the Commission's shortcomings and ask about action that must be taken. While the literature provides a range of factors that impact evaluation use, our analysis showed that the variances in the questions about the follow-up of evaluations can be explained best by the political conflict between the EP and the Commission during the legislative stage. We expected this to be indicative of the *nature* of the risk of the Commission shirking away from tasks delegated to it by the EP. In line with our expectation, this factor increases the chances that evaluations are used by MEPs.

While the *significance* of a policy was not found to have a significant effect, we believe that political factors are most important for evaluation use for accountability purposes. Both the fact that our rationalistic factors had no impact on evaluation use - especially as we remained close to the meaning of these variables in our measurement - and our outlier analysis pointed in the direction of political explanations. Identifying which other political factors are important for explaining evaluation use for accountability purposes by MEPs will require further research. Given the forward-looking use of evaluations, it is advised, as suggested earlier, to also turn to the literature on agenda-setting and policy change. In this respect, a first avenue for further research on the impact of evaluations would be to investigate how the Commission responds to EPQs; this would be a logical step to trace the impact of EPL evaluations at the political level.

Notes

[1] This article was part of a symposium on 'Accountability in the Post-Lisbon European Union' directed by Gijs Jan Brandsma, Eva Heidbreder and Ellen Mastenbroek.
[2] These evaluation can be carried out by the Commission or by the member states.
[3] This document is a good example of a negotiated agreement between European actors that has mostly served to improve the position of the EP without having to change the European treaties (Brandsma 2013a, 6).

[4] Annex C of the guide (European Commission, 2004: 75) further describes what this information should be about: an ex-post evaluation should assess causality and make clear if the criteria of effectiveness, efficiency, relevance and sustainability have been met in a certain intervention. The Commission recommends external evaluations for the purpose of accountability, as they would be more independent (p. 14). If accountability is the aim, the judgment should be entirely with the evaluator and not with the steering group (p. 89).
[5] Højlund (2014) takes the article by Johnson et al. (2009) as a second starting point. This article confirms the framework of Cousins and Leithwood, but adds the impact of participatory evaluation techniques to evaluation use. As this review article explicitly excluded accountability studies (2009: 380) it is not used here as a starting point, although we do refer to the article at times.
[6] The available data does not allow for testing all these factors in a single model. We therefore selected the factors we expected to have a particularly strong impact on the use of evaluations for accountability purposes.
[7] In the past, Commission officials have admitted that too much involvement of policy implementers in evaluation studies can affect the objectivity of the evaluation and its conclusions (EPEC, 2005: 41).
[8] Cousins and Leithwood (1986) also argue that the findings of an evaluation matter. In our view, however, there will not be a difference between evaluations with positive or negative findings, as both may lead to parliamentary questions, depending on the preference of an MEPs regarding a particular issue. Another aspect of evaluation implementation we do not take on board is *timeliness* of the report, as this cannot be reliable measured.
[9] Please note that personal characteristics, commitment to evaluation and information needs of MEPs (Johnson et al., 2009: 356) are not relevant for our purposes, because the unit of analysis are evaluations, not MEPs. We also do not include the availability of competing information, which cannot be reliably measured quantitatively.
[10] The original dataset contains 216 evaluations. During the data collection for this study four evaluations were found and added to the original dataset.
[11] http://ec.europa.eu/smart-regulation/evaluation/search/search.do
[12] http://eur-lex.europa.eu/homepage.html?locale=en
[13] http://www.europarl.europa.eu/plenary/en/parliamentary-questions.html#sidesForm. Last accessed 10 January 2015.
[14] We used a Boolean search term that was tailored for every single evaluation. The search term was: '(evaluation OR study OR report OR review OR assessment) AND ([legislation number] OR [key-words from evaluation report]).' Searching before 1999 would be useless as our dataset of evaluations starts in 2000.
[15] In the cases with a positive number >2, the predicted chances of an evaluation being followed up were slight, whereas in reality EPQs were asked. In cases with a negative number <2, the predicted chances were high, but were not referred to.

References

Ackrill R and Kay A (2011) Multiple streams in EU policy-making: the case of the 2005 sugar reform. *Journal of European Public Policy* 18(1): 72-89.

Bachtler J and Wren C (2006) The evaluation of EU cohesion policy: research questions and policy changes. *Regional studies* 40(2): 143-153.

Bauer MW (2006) Co-managing program implementation: conceptualizing the European Commission's role in policy execution. *Journal of European Public Policy* 13(5): 717-735.

Baumgartner FR, Jones BD and Mortensen PB (2014) Punctuated Equilibrium Theory: Explaining Stability and Change in Public Policymaking. In: Sabatier PA and Weible CM (eds) *Theories of the Policy Process (3rd edition)*. Boulder, US: Westview, pp. 59-103.

Benjamin LM (2008) Evaluator's role in accountability relationships. *Evaluation* 14(3): 323-343.

Blom-Hansen J (2005) Principals, agents, and the implementation of EU cohesion policy. *Journal of European Public Policy* 12(4): 624-648.

Bovens M (2010) Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics* 33(5): 946-967.

Bovens M, 't Hart P and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford Handbook of Public Policy.* Oxford: University Press, pp. 319-335.

Brandsma GJ (2012) The effect of information on oversight: the European Parliament's response to increasing information on comitology decision-making. *International Review of Administrative Sciences* 78(1): 74-92.

Brandsma GJ (2013a) Bending the rules: Arrangements for sharing technical and political information between the EU institutions. In: Ripoll Servent A and Busby A (eds) Agency and influence inside the EU institutions. *European Integration online Papers (EIoP)* 17(1), Article 8, http://eiop.or.at/eiop/texte/2013-008a.htm, pp. 1-22.

Brandsma GJ (2013b) Quantitative Research into Accountability. In: Bovens M, Goodin R and

Schillemans T (eds) *Oxford Handbook of Public Accountability*. Oxford: University Press, pp. 143-158.

Bussmann W (2010) Evaluation of legislation: skating on thin ice. *Evaluation* 16(3): 279-293.

Cobb RW, Ross JK and Ross MH (1976) Agenda Building as a Comparative Political Process. *American Political Science Review* 70(1): 126-138.

Cooksy LJ and Caracelli VJ (2005) Quality, Context, and Use Issues in Achieving the Goals of Meta evaluation. *American Journal of Evaluation* 26(1): 31-42.

Corbett RG, Jacobs FB and Schackleton M (2011) *The European Parliament (8th edition)*. London: Harper.

Cousins JB and Leithwood KA (1986) Current empirical research on evaluation utilization. *Review of Educational Research* 56(3): 331-365.

Curtin D (2007) Holding (Quasi-) Autonomous EU Administrative Actors to Public Account. *European Law Journal* 13(4): 523-541.

Curtin DM (2009) *Executive power of the European Union: Law, practices and the living constitution*. Oxford: University Press.

Curtin DM, Mair M and Papadopoulos Y (2010) Positioning Accountability in European Governance: An Introduction. *West European Politics* 33(5): 929-945.

Datta LE (2006) The Practice of Evaluation Challenges and New Directions. In: Shaw F, Greene JC and Mark MM (eds) *The Sage Handbook of Evaluation*. Thousand Oaks, CA: Sage, pp. 419-438.

EPEC (European Policy Evaluation Consortium) (2005) *Study on the use of evaluation results in the Commission: Final report*. Dossier no. 1: Synthesis report and annexes. Brussels: European Policy Evaluation Consortium.

European Commission (2001) *European governance: A white paper [COM(2001)428]*. Brussels: European Commission.

European Commission (2004) *Evaluating EU activities: A practical guide for the Commission services*. Brussels: European Commission.

European Commission (2007) *Responding to Strategic Needs: Reinforcing the use of*

evaluation [SEC(2007)213]. Brussels: European Commission.

European Commission (2010a) *Multi-annual overview (2002-2009) of evaluations and impact assessments.* Secretariat-general, May 2010. Found 1 December, 2011, at http://ec.europa.eu/dgs/secretariat_general/evaluation/docs/multiannual_overview_en.pdf

European Commission (2010b) *Commission work programmes*. Found 24 April, 2012, at http://ec.europa.eu/atwork/programmes/index_en.htm

European Commission (2013) *Regulatory Fitness and Performance (REFIT): Results and next steps [COM(2013)685 final]*. Brussels: European Commission.

European Court of Auditors (2010) *Impact assessments in the EU institutions: Do they support decision-making? [Special report no. 3]*. Luxembourg: European Court of Auditors.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU Internal Market and Services law. *Evaluation* 18(4): 477-499.

Fleisscher DN and Christie CA (2009) Evaluation Use: Results From a Survey of U.S. American Evaluation Association Members. *American Journal of Evaluation* 30(2): 158-175.

Forss K and Carlsson J (1997) The quest for quality – or can evaluation findings be trusted? *Evaluation* 3(4): 481-501.

Franchino F (2000) Control of the Commission's Executive Functions Uncertainty, Conflict and Decision Rules. *European Union Politics* 1(1): 63-92.

Højlund S (2014) Evaluation use in evaluation systems - the case of the European Commission. *Evaluation* 20(4): 428-446.

Johnson K, Greenseid LO, Toal SA, King JA, Lawrenz F and Volkov B (2009) Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation* 30(3): 377-410.

Kingdon JW (1995) *Agendas, alternatives and public policies (2nd edition)*. New York: Longman.

Laffan B (1999) Becoming a 'living institution': The evolution of the European court of

auditors. *Journal of Common Market Studies* 37(2): 251-268.

Lehtonen M (2005) OECD Environmental Performance Review Program. *Evaluation* 11(2): 169-188.

Levy R (2001) EU Programme Management 1977-96: A Performance Indicators Analysis. *Public Administration* 79(2): 423-444.

Linter P and Vaccari B (2005) The European Parliament's Right of Scrutiny over Commission Implementing Acts: a Real Parliamentary Control? *EIPASCOPE* 1: 15-25.

Long JS and Freese J (2006) *Regression Models for Categorical Dependent Variables Using Stata (2nd edition)*. College Station, Texas: Stata press.

Mastenbroek E, Van Voorst S and Meuwese ACM (2016) Towards more effective problem-solving? Analyzing the ex-post evaluation of European Union legislation. *Journal of European Public Policy* 23(9): 1329-1348.

Nugent N (2010) *The Government and Politics of the European Union (7th edition)*. Palgrave Macmillan.

Patton MQ (2008) *Utilization Focused Evaluation (4th edition)*. Thousand Oaks, CA: Sage.

Pollack MA (1997) Delegation, agency, and agenda setting in the European Community. *International Organization* 51(1): 99-134.

Poptcheva EM (2013) *Library Briefing. Policy and legislative evaluation in the EU*. Brussels: European Parliament.

Princen S (2013) Punctuated equilibrium theory and the European Union. *Journal of European Public Policy* 20(6): 854-870.

Princen S and Rhinard M (2006) Crashing and creeping: agenda-setting dynamics in the European Union. *Journal of European Public Policy* 13(7): 1119-1132.

Proksch SO and Slapin JB (2010) Parliamentary Questions and Oversight in the European Union. *European Journal of Political Research* 50(1): 53-79.

Rossi PH, Lipsey MW and Freeman HE (2004) *Evaluation: A systematic approach (7th edition).* Thousand Oaks, CA: Sage.

Stame N (2008) The European project, federalism and evaluation. *Evaluation* 14(2): 117-140.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 67-85.

Stufflebeam DL and Shinkfield AJ (2007) The nature of program evaluation theory. In: Stufflebeam DL and Shinkfield AJ (eds) *Evaluation. Theory, Models and Applications*. San Francisco, CA: Jossey-Bass, pp. 57-79.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation*. New Brunswick: Transaction, pp. 407-424.

Toulemonde J (2006) Appropriation des résultats de l'évaluation: leçons de la pratique en Région Limousin. In: Genard JL, Jacob SB and Varone F (eds) *L'évaluation au niveau regional.* Brussels: Peter Lang, pp. 131-142.

Vedung E (1997) *Public policy and program evaluation*. New Brunswick: Transaction.

Versluis E, Van Keulen M and Stephenson P (2011) *Analyzing the European Union Policy Process*. Houndmills, Basingstoke: Palgrave MacMillan.

Warntjen A (2012) Measuring salience in EU legislative politics. *European Union Politics* 13(1): 168-182.

Weiss CH (1993) Where politics and evaluation research meet. *American Journal of Evaluation* 14(1): 93-106.

Weiss CH, Murphy-Graham E and Birkeland S (2005) An Alternate Route to Policy Influence: How Evaluations Affect D.A.R.E. *American Journal of Evaluation* 26(1): 12-30.

Wille AC (2010) Political-Bureaucratic Accountability in the EU Commission: Modernising the Executive. *West European Politics* 33(5): 1093-1116.

Wille AC (2012) The Politicization of the EU Commission: Democratic Control and the Dynamics of Executive Selection. *International Review of Administrative Sciences* 78(3): 383-402.

Chapter 9: Discussion and conclusion

Stijn van Voorst

1. Research aims

This dissertation started with the aim of assessing to what extent the European Commission's system for ex-post legislative (EPL) evaluations is fit to contribute to learning and accountability. As was illustrated by the seed and plant propagating material evaluation that was described in the introduction (Arcadia International et al., 2008), as well as by other cases mentioned throughout this dissertation, there are many problems that may plague EPL evaluations and therefore make them unfit to achieve their purposes. Some evaluations are initiated just two or three years after legislation enters into force, which may be too soon to fully assess its effects on society; in other cases evaluations are only initiated after decades or not at all. Whereas some evaluations are of high quality, others lack even a basic description of their research aims and methodologies. Some EPL evaluations are used extensively by the Commission when drafting legislative proposals; others produce findings that receive no serious attention from any EU institution.

Chapter 2 of this dissertation provided a general overview of the Commission's system for EPL evaluations. The six subsequent chapters sough to describe and explain the variation that was found among these evaluations in more detail. Specifically, these chapters answered the three central research questions that are described below.

The first research question was how the variance in the initiation of the Commission's EPL evaluations can be explained (chapter 4). The underlying assumption of this question is that a system of EPL evaluations can only contribute to learning and accountability if it meets the requirement of *systematic initiation*: all major legislation should be evaluated periodically and any omissions in this regard should be explained transparently. Although EPL evaluations may

lead to the improvement of specific laws even if this requirement is not met, in that case they will not enhance legislative quality as a whole (OECD, 2015: 120). If the Commission conducts EPL evaluations selectively it could also create the impression that it decides what legislation to evaluate based on political motives (Radaelli and Meuwese, 2010: 146). Such a reputation could harm the credibility of all its subsequent evaluations.

The second research question of this dissertation was how variance in the *quality* of the Commission's EPL evaluations can be explained (chapter 5). The underlying assumption of this question was that because evaluations are a form of applied research (Pawson and Tilley, 1997: p. xiii), they must meet standards of methodological rigor to contribute to learning (OECD, 2015: 121; Mayne and Schwartz, 2005: 1). If the Commission's EPL evaluations are not valid and reliable, any decisions that take these evaluations into account are based on misleading information. Furthermore, a lack of methodological quality can create the perception among decision-makers that evaluation findings misrepresent reality, which makes it less likely that such findings will be used for learning in the future (Mayne and Schwartz, 2005: 6).

The third research question addressed by this dissertation was how the variance in the *use* of the Commission's EPL evaluations can be explained (chapter 6-8). The assumption underpinning this question is that evaluations only contribute to aims like learning and accountability if their results are seriously considered by decision-makers and other political actors (Klein Haarhuis, 2016: 13; Mayne, 2014). In other words, even if the Commission's EPL evaluations meet the standards of systematic initiation and high quality described above, their results still need to feed into the legislative process to lead to anything but theoretical knowledge (Højlund, 2014).

To explain the variation in the initiation, quality and use of the Commission's EPL evaluations, this dissertation assessed the effects of a broad range of independent variables. Chapter 3 focused on one of these factors: the variation in evaluation capacity of the Commission's directorates-general (DGs). The data presented in this chapter were used for the quantitative analyses presented in several of the other chapters.

The second section of this conclusion answers the three research questions described above, by summarizing the main findings of the preceding chapters. Section 3 discusses the implications of these results for existing theories about the European Commission and policy evaluations. The fourth section of this conclusion places the findings in a broader context by comparing the Commission's system for EPL evaluations to various national evaluation systems. Section 5 discusses three limitations of the dissertation and describes an agenda for future research; in section 6 a number of practical recommendations for the EU institutions are provided. This conclusion ends with some overall reflections about the Commission's system for EPL evaluations in section 7.

2. Answers to the research questions

Table 1 summarizes the descriptive and explanatory results of the chapters of this dissertation. Both this table and the text of this section only present the most recent data related to the three research questions specified above. The results about initiation and quality from chapter 2 are omitted because more up-to-date versions of these findings were discussed in chapter 4 and 5 of this dissertation; the data from chapter 3 are not mentioned separately either because their main purpose was to measure independent variables that were used in other chapters.

Starting with the initiation of EPL evaluations (research question 1), chapter 4 of this dissertation showed that out of the 277 major European regulations and directives published between 2000 and 2004, only 116 (about 42%) had been evaluated at the end of 2014. Although it is theoretically possible that some of this legislation was evaluated after 2014, the chances of this are small because almost no legislation is evaluated after more than ten years.[1] Therefore, it can be deduced that more than half of the major pieces of EU legislation from 2000-2004 have never been evaluated. This shows that the Commission only partly meets the requirement of systematic initiation.

Table 1: overview of research results

Chapter	Topic	Main descriptive results	Significant explanations
4	Initiation of EPL evaluations	116/277 (42%) major EU directives and regulations from 2000-2004 have been evaluated by the Commission.	1. Type of legislation 2. Legislative complexity 3. Evaluation clause 4. Evaluation capacity
5	Quality of EPL evaluations	115/153 (75%) EPL evaluations that assess effectiveness meet five or more out of nine criteria for quality.	Type of evaluator
6	Use of EPL evaluations by IAs	33/51 (65%) IAs for which an EPL evaluation was available used it.	Timeliness
6	Use of IAs by EPL evaluations	10/60 (17%) EPL evaluations for which an IA was available used it.	None found
7	Instrumental use of EPL evaluations by the Commission	Results from EPL evaluations are entirely followed up in some cases, partly followed up in other cases and not followed up in other cases.	Salience of the evaluation's policy field in the eyes of the current Commission.
8	Accountability use of EPL evaluations by the European Parliament	49/220 (22%) EPL evaluations were referred to in EP questions. 34/220 (16%) EPL evaluations were used to steer the Commission's behaviour.	Level of conflict between the EP and the Commission during the legislative stage.

As was explained in chapter 4, one of the reasons why systematic initiation is important is that the credibility of an evaluation system decreases if evaluations appear to be selectively conducted for political reasons. Therefore, the low initiation rate presented above raises the question whether or not the Commission decides what legislation to evaluate based on its own political interests. To study this question empirically, chapter 4 tested the hypothesis that the Commission is less likely to initiate an EPL evaluation when the risk that its competences could be reduced is higher (as measured by proxies concerning the legislative procedures used in the EP and the Council). This hypothesis was rejected, which means that no evidence was found for the presence of this type of political considerations.

Four other factors, however, do affect the variance in the initiation of EPL evaluations by the Commission. Firstly, the type of legislation matters: directives are more likely to be evaluated than regulations. Secondly, the chances that a piece of legislation is evaluated increase with its complexity. Both of these explanations suggest that the Commission may prioritize evaluating legislation that grants more freedom to the member states, because for such legislation the risk of non-compliance is higher. In other words, EPL evaluations may partly be initiated by the Commission to make its task of enforcing EU legislation easier.

A third significant explanation for the variance in the initiation of EPL evaluations by the Commission is the presence of evaluation clauses. Legislation containing a provision that requires it to be evaluated within a given number of years is significantly more likely to be evaluated than legislation without such a provision. However, the Commission does not always meet the deadlines or fulfil the other demands specified by such clauses. A fourth significant explanation for the initiation of EPL evaluations is the evaluation capacity of the responsible DG. DGs with a specialized unit for ex-post evaluation and/or specific guidelines for EPL evaluations evaluate a significantly higher proportion of their legislation than other DGs. Therefore, the presence of sufficient capacity to evaluate appears to make it easier for DGs to meet the requirement of systematic initiation, although it is also possible that DGs that initiate more EPL evaluations build more capacity to be able to support such evaluations.

Concerning the topic of quality (research question 2), chapter 5 of this dissertation showed that the vast majority (75%) of the Commission's 153 evaluations that address legislative effectiveness meet five or more out of criteria on which they were assessed. Some 76% of these evaluations use both stakeholder input and other research methods, which suggests that their methodology is built on a robust comparison of different data sources.

However, the evaluations perform less well regarding other aspects of quality. Whereas almost all reports (89%) have a well-defined scope in the sense of clearly specified research questions, less than 40% of them go beyond this by also describing the intervention logic of the legislation that they evaluate. Between 40% and 70% of the EPL evaluations meet criteria like the presence of a clear operationalization (internal validity), a clear country selection and a clear case selection (external validity) and the presence of substantiated conclusions. By far the worst aspect of the evaluations' quality is their replicability: only 31% of the reports contained or referred to all the material that would be required to repeat the underlying research (like interview guides and lists of respondents).

How can this variance in quality be explained? The explanatory analysis of chapter 5 revealed the type of evaluator as the key determinant: EPL evaluations conducted by external consultants are of significantly higher quality than evaluations conducted internally by the Commission. This suggests that the technical expertise of external parties is a crucial asset when it comes to properly evaluating EU legislation. The evaluation capacity of the Commission's DGs, the complexity of the evaluated legislation and various political conditions were found to have no effect on the variance in quality.

As for the use of the Commission's EPL evaluations (research question 3), chapter 6 of this dissertation revealed that out of the 51 impact assessments (IAs) published between 2003 and 2014 for which a prior EPL evaluation was available, 33 cases (65%) made use of at least some information from that EPL evaluation. This shows that the use of the Commission's EPL evaluations for improving legislative proposals (a type of learning) occurs relatively frequently when it is possible. However, in many cases the IA only included some basic references to the EPL evaluation: in-depth use was relatively rare. The only significant explanation for the use of

EPL evaluations by IAs turned out to be timeliness: if the EPL evaluation is not available at least a year before an IA is published, it is almost never used.

Chapter 7 of this dissertation also discussed the use of EPL evaluations for improving legislative proposals, this time with a focus on political explanations. Based on in-depth case studies of the seed law evaluation mentioned above and two other evaluations concerning consumer protection and animal welfare, this chapter showed that opposition from key stakeholders like the European Parliament (EP), the Council and major interest groups does not prevent the Commission from using the results of EPL evaluations. However, the salience of the evaluated policy field in the eyes of the Commission did turn out to be a necessary condition for use. In other words, EPL evaluations dealing with policy fields unrelated to the political priorities of the current Commission are unlikely to affect legislative proposals, whereas EPL evaluations addressing policy fields that fit with these political priorities can influence the exact content of plans for legislative amendments.

Whereas chapter 6 and 7 of this dissertation focused on the use of the Commission's EPL evaluations for policy improvement, chapter 8 discussed their use for a second purpose: enhancing accountability towards the legislature (Højlund, 2014: 444; Vedung, 1997: 102-108). The chapter showed that the EP rarely uses EPL evaluations to hold the Commission accountable: just 49 out of the 220 EPL evaluations studied in this chapter (22%) were referred to in the EP's questions to the Commission at least once. In 34 out of 220 cases (16%), such questions served to steer the Commission's behaviour. These numbers suggest that the use of EPL evaluations for accountability purposes by the EP is uncommon - especially considering the fact that the results of most of these evaluations are summarised in official communications from the Commission that are received and discussed by parliamentary committees. However, not all EP questions explicitly mention their sources, so it is possible that more of these questions are based on EPL evaluations than the data suggest.

Variation in the extent to which EPL evaluations were used in EP questions turned out to be best explained by the degree of conflict about the evaluated legislation between the Commission and the EP during the legislative process. The higher the level of conflict, the higher

the odds that an evaluation was referred to in EP questions. A theoretical explanation for this relation is the nature of the risk faced by the EP: when members of the EP perceive that the Commission may deviate from its wishes when implementing legislation, they become more likely to closely scrutinize the institution. For such scrutiny, EPL evaluations are a potential source of information. By contrast, technical factors like the quality and relevance of the EPL evaluations were found to have no effect on the use of these evaluations in EP questions. The salience of the evaluated legislation in the eyes of the EP also provided no significant explanation.

3. Theoretical implications

The chapters of this dissertation have discussed the effects of various political and technical variables on the initiation, quality and use of the Commission's EPL evaluations. Whereas the explanatory power of these individual factors has been summarised above, the broader implications of the findings have not yet been addressed. This section therefore places the results of this dissertation in the context of some general theories about the interests of the European Commission and about evaluation systems. The first subsection below discusses the findings in relation to existing theories about the Commission's interests in protecting its competences and encouraging European integration. Subsection two focuses on other theoretical implications related to political interests, which were found inductively based on the case studies about evaluation use. In the third subsection the theoretical implications of the findings about technical explanations are discussed.

Political explanations: protection of competences and encouraging EU integration

Since EPL evaluations assess the functioning of government policies, they inherently take place in a political environment (Bovens et al., 2008; Weiss, 1993). Due to this political setting, evaluation results usually benefit some actors and put others at a disadvantage, for example because they allocate praise or blame for certain policy outcomes or (re)open debates about sensitive issues (Bovens et al., 2008: 320; Schwartz, 1998: 295; Weiss, 1993: 95-98). Political

factors can therefore be expected to affect the Commission's EPL evaluations and were taken into account throughout this dissertation. Such factors have been defined as interests that actors have in (not) conducting evaluation-related activities like initiating an evaluation, investing in its quality and using its results.

To discuss the effects of the Commission's political interests on its evaluation activities, these interests must first be specified. One commonly used branch of theory that addresses the motives of political actors is public choice. According to this theoretical framework, civil servants mainly aim to maximize their budgets (e.g. Niskanen, 1971), whereas politicians seek to maximize votes (e.g. Dunleavy, 1991). However, as various scholars in the field of EU governance have argued, these views cannot be directly applied to the Commission, as that institution is not directly elected and operates on the basis of a fixed budget that is mostly spent on agriculture and regional development (Tallberg, 2003: 28; Majone, 1996: 65). Furthermore, in the context of this dissertation budgetary maximization is not very important because most EPL evaluations do not concern financial incentives.

Rather than presenting the Commission as a maximizer of voters or budgets, the literature that views the Commission as a rational actor generally argues that the institution seeks to protect and/or expand its competences (Majone, 1996: 65; Nugent and Rhinard, 2016: 1201; Pollack, 2008: 9; Tallberg, 2003: 28). Hartlapp et al. (2014: 1-14) call this the perspective of the Commission as a 'competence seeker', in contrast to viewing the institution as a 'policy seeker' or a technocratic institution. Theoretically, the Commission has an incentive to maximize its competences because this makes it easier to pursue its policy aims (Majone, 1996: 65; Pollack, 2008: 9). Furthermore, the Commission frequently faces threats and opportunities related to the scope of its powers upon which it needs to act. For example, during the last decade the Commission has succeeded in strengthening its role in budgetary oversight, while its ability to produce secondary legislation in certain policy fields has been reduced due to increased oversight by the EP (Nugent and Rhinard, 2016).

Especially when their findings are negative, EPL evaluations may open discussions about the competences of the actors that implement legislation. Such actors may therefore seek to

avoid evaluations to prevent criticism ('blame avoidance') (Van Thiel, 2016). Therefore, chapter 4 of this dissertation tested the hypothesis that the chances of an EPL evaluation being initiated by the Commission are lower for policies for which there is a higher risk that the evaluation's results could lead to legislative amendments that reduce the institution's competences. Furthermore, chapter 5 tested the hypothesis that the quality of the Commission's EPL evaluations is lower under such circumstances. The quantitative analyses presented in these chapters falsified both of these hypotheses. In other words, no evidence was found in line with the theory that the Commission primarily seeks to protect and/or expand its own competences.

Chapter 7 of this dissertation addressed the question whether or not the Commission's use of EPL evaluations can be explained by the extent to which the results of these reports are in line with its pre-existing preferences about the evaluated legislation. In two out of the three analysed cases, the Commission held the pre-existing belief that its own competences should be increased. However, the qualitative data presented in this chapter suggested that these preferences had no significant impact on the Commission's subsequent decisions about the use of the evaluations. In other words, the case studies also provided no evidence in line with the hypothesis that the Commission would be primarily motivated by the protection or expansion of its competences.

Besides the issue of competence maximization, a part of the literature about the Commission argues that the institution is (also) driven by its intention to enhance European integration (e.g. Nugent and Rhinard, 2016: 1208; Pollack, 2008: 9; Tallberg, 2003: 28). This issue is partly related to the increase of competences, as more European integration can lead to increased powers for the Commission, but it can also be a motivation of its own when the institution considers itself to have an inherent responsibility to promote European integration.

One task related to the promotion of European integration is the Commission's role as the 'guardian of the treaties': an enforcer of member state compliance with EU legislation (Steunenberg, 2010: 359; Tallberg, 2003: 28). As has been argued at various points in this dissertation, EPL evaluations are one potential tool to fulfil this task, since they may provide the Commission with information about how national administrations implement EU legislation.

This purpose is not mentioned as one the official functions of EPL evaluations in the Commission's better regulation documents (2013: 2; 2015: 259), which focus on the aims of learning and the Commission's own accountability towards other institutions. However, the need to enforce legislation can be expected to play a role in the practice of EPL evaluations if the Commission is indeed motivated by its interest in encouraging integration.

Chapter 4 of this dissertation therefore tested the expectation that the Commission is more likely to evaluate a piece of legislation when the resulting EPL evaluation is more useful for enforcement purposes. The results showed that the Commission indeed prioritizes evaluating policies for which the chances of non-compliance by the member states are higher, such as directives and complex legislation. These findings suggest that the enforcement of EU legislation indeed plays a role in the Commission's initiation of EPL evaluations.

To summarize, whereas the results of this dissertation are not in line with the theoretical image of the Commission as a 'competence seeker', they do present some evidence that the institution aims to enforce European integration when taking decisions about EPL evaluations.

Political explanation: the intention to reduce legislative output

In addition to the findings discussed above, the case studies conducted for this dissertation inductively revealed another factor that affects the Commission's evaluation-related activities: its intention to reduce its legislative output. This explanation is closely related to recent developments in the Commission's activities.

When it entered into office at the end of 2014, the Juncker Commission announced its intention to focus its efforts on a limited number of policy fields (like the economy, migration and human rights - see Juncker, 2014). As a result, the Commission (2016: 2-3) now launches only dozens of legislative proposals annually, whereas this number was in the hundreds in the past. One alleged reason for this shift was to combat Euroscepticism by convincing citizens and companies that the EU can be 'be small on small issues' (i.e. not to overregulate society) (European Commission, 2016: 2; Juncker, 2014). The current Commission also faces more

pressure from the European Council to restrict its legislative activities than was the case for its predecessors (Nugent and Rhinard, 2016: 1205-1206).

Chapter 7 of this dissertation suggested that this drive to reduce legislative output strongly affects the Commission's use of EPL evaluations. An in-depth analysis of three high-quality EPL evaluations showed that the recommendations of such reports are unlikely to be implemented when they concern policy fields that are no priorities of the current Commission (like plant health and animal welfare). In other words, the salience of the policy field to which an EPL evaluation belongs appear to be a necessary condition for the use of its results. This imposes a political constraint on the Commission's use of EPL evaluations: the conclusions of such reports may be used when drafting legislative proposals, but only if the political top of the Commission wishes to take action in that policy field to begin with.

These results about the use of EPL evaluations have some broader implications for the Commission's better regulation agenda. On the one hand, a key element of this agenda is to strengthen evidence-based policy: the legislation proposed by the Commission should be based on objective information, for which EPL evaluations are one potential source (European Commission, 2013: 5; 2015: 253; 2016: 2, 7). On the other hand, the better regulation agenda also implies that the EU should be careful not to produce too much new legislation, as the burdens that such rules impose on society should remain as small as possible (2012: 4-5; 2015: 254; 2016: 6-7). These two aims of the better regulation agenda potentially clash in the case of EPL evaluations. Most of these evaluations recommend some changes to legislative texts to improve them, which usually requires the Commission to propose amendments. To follow-up on such recommendations is in line with evidence-based policy, but can go against the aim of reduced legislative output. Chapter 7 of this dissertation suggested that the Commission currently prioritizes the second aim over the first.

This tension between evidence-based policy and deregulation is not unique to the Commission: better regulation agendas in general have multiple aims that may conflict with each other and shift over time (Bunea and Ibenskas, 2017: 593; Radaelli, 2007: 192-193). However, when compared to most national administrations the Commission faces more

pressure to reduce its legislative output (Nugent and Rhinard, 2016: 1205-1206). This pressure may limit the institution's use of EPL evaluations and should therefore be taken into account to fully understand the dynamics of evidence-based policy within the Commission.

Technical explanations

Besides political explanations, this dissertation also discussed the effect of various technical variables on the initiation, quality and use of the Commission's EPL evaluations. In the context of this research, 'technical' factors refer to practical prerequisites that affect the functioning of evaluations. These factors are rooted in a rational and apolitical perspective on evaluations, which means that evaluations are seen as tools that can produce objective information when they follow the right procedures and involve the right actors (Bovens et al., 2008: 325). Unlike the political influences discussed above, the technical factors studied in this dissertation were mostly derived from general literature about policy evaluations (e.g. Cooksy and Mark, 2012; Nielsen et al., 2011; Rossi et al., 2004).

The exact technical variables that were studied in this dissertation varied somewhat from chapter to chapter, depending on the topic at hand. The technical variable that was most consistently discussed is evaluation capacity: the presence of organizational resources that are meant to support evaluated-related activities (Nielsen et al., 2011: 325; Stockdill et al., 2002: 14). Existing theories predict that evaluation capacity positively affects the initiation, quality (Cooksy and Mark, 2012: 81) and use (Stockdill et al., 2002: 14) of evaluations, as it allows for more investments in every stage of an evaluation process, from data-collection (Nielsen et al., 2011: 327) to the presentation of a final report (Rossi et al., 2004: 414).

Chapter 3 of this dissertation described the capacity of the Commission's DGs to conduct EPL evaluations. These findings were subsequently used in chapter 4 and 5 to test the hypotheses that the variation in the initiation and quality of EPL evaluations between DGs depends on the variation in their evaluation capacity. The results confirmed that DGs with more capacity (as measured by them having an evaluation unit and evaluation guidelines) indeed

evaluate a higher proportion of their legislation than other DGs. However, the findings also showed that the EPL evaluations of these DGs are not of significantly better quality.

The research conducted for this dissertation found that the Commission's EPL evaluations are affected by various other technical factors as well, such as the presence of evaluation clauses and the timeliness of the evaluations. These variables have been listed per topic in Table 1. Overall, the presented findings are in line with the theoretical view that high-quality policy evaluations are more likely to be initiated and used when a number of practical prerequisites are met. In other words, the practice of EPL evaluations in the Commission is affected not only by political conditions, but by technical factors as well.

4. Comparison with other evaluation systems

The findings presented above show that the Commission's EPL evaluations are far from perfect in terms of their initiation, quality and use. However, so far this conclusion only discussed these evaluations in absolute terms: their relative merits when placed next to evaluations from other political systems remain unclear. Therefore, to properly contextualize the results of this dissertation, this section compares the Commission's system for EPL evaluations to some national systems for such evaluations.

For this assessment the Commission will be compared to the member countries of the OECD. Such a comparison is worthwhile because the OECD (2015: 122) actively promotes the institutionalization of EPL evaluations, which makes its members relatively likely to have systems for such evaluations in place. In 2015 the OECD published its most recent 'regulatory policy outlook', which assessed (among other topics) to what extent its 34 member countries and the European Commission possess a system for EPL evaluations. This report serves as the basis for the comparisons made in this section.

Overall, the OECD (2015) is critical of the practice of EPL evaluation in its member states. It states that most of these countries do not have structures in places that support such evaluations (p. 129), that most national EPL evaluations do not fully assess the economic and social impact of legislation (p. 130-131), and that the quality of these evaluations is insufficiently

supervised (p. 132-133). Out of the 35 political entities studied in the OECD's report, only seven[2] had systematically initiated EPL evaluations that went beyond administrative burden calculations during the years before 2015 (OECD, 2015: 16). The European Commission is among this small group (OECD, 2015: 30, 158-159).

By implication, the remaining 28 countries do not systematically conduct EPL evaluations. In most of these countries some EPL evaluations take place on an ad hoc basis, but formal procedures for their initiation, quality and use are missing, unclear or rarely applied in practice (OECD, 2015: 142-211). For example, although the French government has produced initiation procedures (National Assembly of France, 2015: 100-101), methodological guidelines (Scientific Council for Evaluation, 1994) and quality assurance systems (National Council of Evaluation, n.d.) for ex-post evaluations, none of these documents and procedures address legislative evaluations specifically. In practice, French EPL evaluations are mostly conducted in the context of ad hoc modernization programmes launched by the government (Government of France, n.d.; OECD, 2015: 162).

In comparison with these countries that have no clear standards for EPL evaluations in place, the Commission's initiation, quality and use of such evaluations is more systematic by default. But how does the Commission perform when compared to the few OECD countries that do have systematic procedures for EPL evaluations? Whereas an in-depth study of the similarities and differences with each of these countries would go beyond the scope of this dissertation, a comparison with one particular example can be made. The Netherlands is suitable for such an assessment, as its systems for EPL evaluations and evaluation clauses have been subject to extensive empirical research (Klein Haarhuis and Niemeijer, 2009; Klein Haarhuis, 2016; Von Meyenfeldt et al., 2017; Veerman et al., 2013). The Netherlands is also comparable to the European Commission in that it started to develop an evaluation system for spending activities in the late 1980s, after which it gradually developed ex-post evaluations for legislation and other activities in the subsequent decades (Leeuw et al., 2009: 90-97).

There are two main types of EPL evaluations in the Netherlands: legislative evaluations and policy reviews.[3] Legislative evaluations, firstly, concern individual laws or groups of laws.

Policy reviews, secondly, evaluate all government actions in a certain policy area - which usually includes legislation (*Regeling Periodiek Evaluatie-onderzoek*, 2014; Klein Haarhuis, 2016: 10). This makes such reviews somewhat comparable to fitness checks at the EU level, which also assess whole policy areas (European Commission, 2015: 254).

Regarding the *initiation* of EPL evaluations in the Netherlands, firstly, it is difficult to find numbers that are entirely comparable to the data about the Commission presented in this dissertation. However, some impressions can be gained by studying the total number of EPL evaluations as compared to the number of legislative proposals in both political systems.

The Dutch Ministry of Justice lists 43 legislative evaluations that were completed by government departments in 2013 and 2014 (Knowledge Centre Legislation and Legal Affairs, n.d.), which is less than the 87 EPL evaluations completed by the Commission's DGs during those years according to the dataset used for this dissertation. Furthermore, a meta-evaluation of policy reviews commissioned by the Ministry of Finance showed that the Dutch government had completed 23 of such reports in early 2017 (Von Meyenfeldt et al., 2017: 5), which is a much less than the 47 fitness checks that had been published by the Commission in 2014 (Smismans, 2015: 14). At first sight this variation cannot be explained by the differences in total legislative activity between the Netherlands and the EU. For example, in 2013 and 2014 the Dutch government sent respectively 229 and 256 bills to parliament (Tweede Kamer der Staten-General, n.d.), whereas these numbers were respectively 66 and 120 for the Commission, which has reduced its legislative activity even further since then (European Commission, 2016: 3). Therefore, the Dutch government also appears to initiate fewer EPL evaluations than the Commission in relative terms.

Furthermore, there appear to be fewer legal obligations to initiate EPL evaluations for the Dutch government than for the Commission. In 2013 only 10-20% of Dutch primary laws included an evaluation clause (Klein Haarhuis, 2016: 11; Veerman et al., 2013), which is significantly less than the 60% (165/277) of major EU laws from 2000-2004 that were found to contain such a clause in chapter 4 of this dissertation. In the absence of an evaluation clause, conducting an EPL evaluation can still be compulsory in the Netherlands based on the annual

budgetary programmes of individual ministries. However, research has shown that such programmes may not be complied with in practice, depending on the priorities of the Dutch cabinet (Klein Haarhuis, 2016: 11).

Regarding the *quality* of Dutch EPL evaluations, secondly, a meta-evaluation of 75 of such evaluations showed that 81% of them contain a clear problem definition and 85% of them apply a combination of research methods (triangulation) (Klein Haarhuis and Niemeijer, 2009: 410). This is a little worse than the Commission's EPL evaluations studied in chapter 5 of this dissertation, of which 89% contain a clear problem definition and 89% use multiple research methods. On the other hand, the Dutch study also showed that 85% of its evaluations clearly define their concepts (Klein Haarhuis and Niemeijer, 2009: 410), which is much more than the 61% of the Commission's EPL evaluations that were found to contain a clear operationalization. Based on this comparison, it appears that Dutch EPL evaluations perform worse than the Commission's EPL evaluations in some respects, but perform better regarding other criteria. However, it should be noted that the data may not be entirely comparable, for example because of differences in the studies' timeframes.

Other meta-evaluations also assessed the quality of Dutch EPL evaluations. For example, a recent report about the evaluation capacity of Dutch ministries suggested that this quality varies greatly between departments (Klein Haarhuis, 2016: 7). Furthermore, a recent meta-evaluation criticized the quality of Dutch policy reviews. Even though these reviews often meet a number of formal requirements concerning their problem definition and independence, most of them do not properly assess the effectiveness and efficiency of policies (Von Meyenfeldt et al., 2017: 6). Earlier research from the Court of Auditors of the Netherlands (2012: 15) also showed that many Dutch ex-post evaluations falsely claim to study effectiveness. The quality standards used by these meta-evaluations are not always quantified and differ somewhat from the ones used in chapter 5. Therefore, their conclusions are more difficult to compare to the results of this dissertation than the findings of Klein Haarhuis and Niemeijer (2009). However, the general picture that these meta-evaluations provide is that the quality of Dutch EPL

evaluations varies considerably from case to case, as is the case with the Commission's EPL evaluations at the EU level.

Concerning the *use* of Dutch EPL evaluations, thirdly, both legislative evaluations and policy reviews must be sent to parliament; policy reviews are expected to be discussed by the Council of Ministers as well (Klein Haarhuis, 2016: 13; Ministry of Finance, 2014: 2). Empirical research shows that even though these procedures are usually complied with, the Dutch government and parliament do not fully utilize most EPL evaluations in practice. Although Dutch politicians and civil servants seem increasingly interested in EPL evaluations as instruments for accountability, their use of such evaluations for learning is much less common (Klein Haarhuis, 2016: 13).

All in all, the comparison presented above reveals that many of the issues with the initiation, quality and use of the Commission's EPL evaluations discussed in this dissertation also occur in the Netherlands. In particular, research shows that the quality and use of Dutch EPL evaluations varies greatly from case to case, similar to the situation at the EU level. Although the initiation of EPL evaluations in the Netherlands is more difficult to assess, the Commission appears to produce more evaluations than the Dutch government annually and evaluation clauses are more common in EU legislation than in Dutch legislation, even when the total legislative activity of both political systems is taken into account.

In conclusion, this section has shown that the Commission's system for EPL evaluations performs well in relative terms. Most OECD countries do not have systematic procedures for EPL evaluations at all, which means that the Commission outperforms them by default. Furthermore, even the few OECD countries that have systematic procedures for EPL evaluations in place appear to face problems concerning their initiation, quality and use, which shows that such issues are not unique to the Commission. Therefore, when considering the shortcomings of the Commission's system that are discussed in this dissertation, it should be remembered that the institution is still ahead of or on par with most national systems for EPL evaluations.

5. Limitations and recommendations for future research

The research that was conducted for this dissertation necessarily has some limitations, which need to be discussed to fully clarify the value of the conclusions that were drawn. Therefore, in this section the three main drawbacks of this dissertation are discussed. Furthermore, for each of these limitations some recommendations are provided about how future research could solve them.

Limitation 1: causal mechanisms

The first limitation concerns chapter 4 and 5 of this dissertation, which discussed the initiation and quality of the Commission's EPL evaluations (research question 1 & 2). These topics were studied with the help of two self-constructed datasets of 277 pieces of major legislation and 313 evaluations. This quantitative approach allowed this dissertation to provide a unique, comprehensive overview of the previously unexplored topic of EPL evaluations in the EU. It also allowed for drawing valid and reliable conclusions about what factors explain the initiation and quality of such evaluations. However, the drawback of this quantitative approach is that the mechanisms that underpin these causal relations remain obscure (George and Bennett, 2005: 21; Lieberman, 2005: 339-340). In other words, it is not always clear *why* certain variables affect the initiation or quality of EPL evaluations. Although the fact that both topics were studied based on hypotheses derived from theory reduces this problem to some extent, it does not entirely remove the issue.

For example, as explained above, the complexity of EU legislation turned out to have a positive relation with its chances of being evaluated. Based on the theoretical literature used in chapter 4 (e.g. Coglianese, 2012: 11; Stame, 2008: 128-130), the explanation for this seems to be that the Commission prioritizes evaluating legislation for which the implementation is likely to be problematic, as more complexity makes it more difficult to establish if member states comply with EU rules. However, it is also possible that the Commission is more likely to evaluate complex legislation due to its REFIT programme, which aims to use EPL evaluations (and other tools) to simplify legislation that is difficult to understand (European Commission, 2012: 3; 2015:

254; 2016: 6-7). In such situations, there are multiple reasons why one variable may affect another. Quantitative research is often unable to take all these explanations into account, as it usually focuses on a limited number of variables that provide the most plausible explanations for a phenomenon (Lieberman, 2005: 435).

Another disadvantage of the quantitative approach is that some variables cannot be observed directly on a large scale. This problem may result in conceptual stretching: the proxies used to measure certain variables may not fully cover the abstract concepts that they are supposed to represent (Lieberman, 2005: 435). For example, in this dissertation the risk that an EPL evaluation could threaten the Commission's competences was measured via two characteristics of the EU's legislative process: the level of EP involvement and the voting procedure in the Council. These indicators assess the circumstances in which the Commission is likely to face a strategic risk when conducting EPL evaluations rather than the risk itself, as the latter is difficult to measure directly. As a result, the way in which this concept was measured may have affected the findings of chapter 4 and 5 to some extent.

To address this issue of conceptual stretching and to fully assess the mechanisms behind the initiation and quality of the Commission's EPL evaluations, future research could use case studies and other qualitative methods (George and Bennett, 2005: 21; Lieberman, 2005: 439-440). One useful approach could be to use the quantitative data that underpin this dissertation (or a later update of this data) to select interesting cases (like typical cases, deviant cases and so forth) for in-depth studies (Lieberman, 2005: 343-344). Semi-structured interviews with civil servants of the Commission and other relevant actors could be used to find the most plausible mechanisms behind causal relations.

Limitation 2: wider impact of the Commission's EPL evaluations

A second limitation of this dissertation relates to chapter 6-8, which discussed the use of EPL evaluations for learning by the Commission and for accountability purposes by the EP (research question 3). In reality, the Commission and the EP are just two of the actors that may use knowledge from such evaluations. The use of EPL evaluations by other actors, like the Council

and the member states, has not been studied in this dissertation, except for some discussion of the influence of these actors on use by the Commission in chapter 7.

This omission is important because theoretically speaking, there are good reasons for these stakeholders to use the Commission's EPL evaluations. The Council can be seen as one chamber of the EU's legislature (next to the EP) and as such it has the task to hold the Commission accountable (Bovens et al., 2010: 29). EPL evaluations can be a source of information to properly fulfil such accountability functions (Højlund, 2014: 429; Luchetta, 2012: 563; Summa and Toulemonde, 2002: 409). The way in which the Council and the member states use the Commission's EPL evaluations (as well as other ex-post evaluations) therefore seems to be an unexplored topic that warrants future research.

Somewhat related to this issue is the fact that this dissertation has focused solely on the use of EPL evaluations for learning and accountability. In reality, evaluations can be used for other aims as well, including strategic and conceptual ones (Vedung, 1997: 102-111). Strategic use refers to an actor using an evaluation to further his own interests. Although chapter 7 and 8 of this dissertation addressed strategic elements of use to some extent, these chapters conceptualized such strategic elements as threats to instrumental and accountability use rather than as a type of use of its own. Conceptual use refers to using evaluations for long-term knowledge building (Mayne, 2014: 3; Vedung, 1997: 110-111), which is a topic that this dissertation has not discussed at all.

To address the point that the long-term consequences of the Commission's EPL evaluations have been ignored so far, future research could focus on their *influence* rather than just their use. The concept of influence refers to the impact of evaluations in the broadest sense of the word (Herbert, 2014: 389-394; Johnson et al., 2009: 378; Henry and Mark, 2003: 310). Influence matters because evaluations that remain unused in the short term may still have an impact in the long run (Herbert, 2010: 390). For example, chapter 7 of this dissertation showed that the evaluation of the EU's animal welfare policy from 2010 remained unused because it did not fit with the priorities of the Juncker Commission. However, the idea of an animal welfare framework law that was proposed by this evaluation has been picked up by various interest

groups and civil servants. Possibly, they can use the evaluation's arguments once again when a new Commission enters into office. In this case and others, evaluations that remained unused in the short term may still have an impact in the long run via the diffusion of ideas and the change of political agendas (Henry and Mark, 2003: 298). It would be interesting to study to what extent and under what circumstances the Commission's EPL evaluations can have such effects.

Limitation 3: timeframe of the research

A third potential limitation of this dissertation concerns its timeframe. As was explained in the preceding chapters, only EPL evaluations published until 2014 have been studied in this research, because not all evaluations completed from 2015 onwards had been published at the time the data collection was completed. The disadvantage of this decision is that it raises the question to what extent the results still represent the current situation in 2018.

The Juncker Commission, which entered into office at the end of 2014, repeatedly stated that is wished to step up the Commission's efforts concerning EPL evaluation (e.g. Juncker, 2014: 6; European Commission, 2016: 2). In particular, it placed its better regulation efforts under the responsibility of the first vice-president (Frans Timmermans) and it published new guidelines concerning the initiation, quality and use of EPL evaluations (among other topics) in 2015 (European Commission, 2015: 253-298). Theoretically, these activities may have altered the practice of EPL evaluations in the Commission to some extent.

Some of the evidence presented in this dissertation suggests that such changes have been limited. As was explained in chapter 3, in early 2015 most evaluation coordinators in the DGs did not think that the new better regulation guidelines would lead to more or better EPL evaluations in the near future. In their view, that would require additional resources, which were not available.

Furthermore, the total number of EPL evaluations that the Commission produces annually does not seem to have increased significantly since 2014. As was explained in the introduction, the better regulation guidelines from 2015 limited the Commission's official definition of EPL evaluations to staff working documents and annexes to IAs only (European

Commission, 2015: 289-290). All external evaluations must now be translated into such documents. The Commission's register[4] contains four of these 'new-style' evaluations from the end of 2015 and 23 of such cases from both 2016 and 2017.[5] These figures cannot be entirely compared to the numbers of evaluations from before 2015 that were presented throughout this dissertation, since at that time the Commission's EPL evaluations still took many different forms. However, these data suggest that at least there has been no large increase in the Commission's output of EPL evaluations.

On the other hand, there is evidence that the Commission's EPL evaluations have changed in two particular ways since 2014. Firstly, the role of the Commission's secretariat-general (SG) has been strengthened. Most evaluation coordinators interviewed for chapter 3 believed that the growing importance of this institution would be the most important change resulting from the better regulation guidelines of 2015. The fact that all EPL evaluations must now be produced as staff working documents gives the SG an increased role in their management, as it always checks such files before their publication (European Commission, 2015: 288). The better regulation guidelines of 2015 also strengthened the SG in other ways, for example by formalizing its role in the steering groups for fitness checks (European Commission, 2015: 263).

Since the SG is closely related to the Commission's president (Kassim, 2018: 788, 794), the former's enhanced role in EPL evaluations could also strengthen the level of political control over these evaluations. Some evidence for such increased political steering was presented in chapter 7, which showed that the Commission's use of EPL evaluations has reduced since 2014 because its political leadership is only willing to propose legislative amendments for topics that fit with its priorities. The animal welfare evaluation that was described in this chapter exemplifies how the SG can block the follow-up of results of EPL evaluations that do not fit with the agenda of the Commission's president.

A second change to the Commission's EPL evaluations since 2014 is the increased role of the Regulatory Scrutiny Board. Since 2015, this semi-independent institution has the official task to assess the quality of all the Commission's evaluations. In practice, it has decided to scrutinize

all IAs and a growing selection of ex-post evaluations, some of which concern legislation (Regulatory Scrutiny Board, 2018: 11). These checks and the associated feedback may have a positive effect on the quality of the evaluations.

In conclusion, although the Commission's EPL evaluations have remained similar in many regards since the data collection for this dissertation was completed, they have changed in the extent to which they are affected by the SG and the Regulatory Scrutiny Board. Therefore, future academic research could focus on the impact that these institutions have on the initiation, quality and use of EPL evaluations. Possibly, the growing role of the SG has increased the impact of political factors on the evaluations and the growing role of the RSB has increased their quality. However, more data would be needed to test these hypotheses.

6. Practical implications

Besides the theoretical issues discussed above, the results of this dissertation also have some implications for the practice of EPL evaluations in the EU. Therefore, based on the findings of the research, this section provides three recommendations for the EU's institutions.

The findings presented above showed that the Commission's EPL evaluations do not always contribute to learning and accountability in practice, due to a combination of political and technical impediments. Nevertheless, the idea that EPL evaluations should contribute to such aims is useful as a normative standard (Sanderson, 2002: 7) and therefore underpins the three recommendations described below. EU legislation potentially affects millions of citizens and companies. In order to make rational decisions about how to improve these effects, institutions like the Commission require systematic evidence about how legislation works in practice, which ex-post evaluations can potentially provide (Böhme, 2002: 99; Sanderson, 2002: 3, 5). Therefore, the Commission's system for EPL evaluations should encourage learning and accountability as much as possible within its political and technical constraints.

The first recommendation relates to the results about evaluation clauses presented in chapter 4. These findings show that the presence of such clauses significantly increases the chances that major EU regulations and directives are evaluated. However, the results also

suggest that the Commission only complies with about 56% of these clauses. Therefore, the presence of an evaluation clause is not a sufficient condition to ensure that an EPL evaluation will be initiated. To further improve the effectiveness of such clauses, it would be useful to have an institution that keeps track of their follow-up.

Such a role could be fulfilled by the civil service of the EP or the Council, as these institutions often insert or amend evaluation clauses in EU legislation to ensure that they remain informed about its implementation (Summa and Toulemonde, 2002: 410). Therefore, they have an interest in checking the extent to which the Commission complies with evaluation clauses. Furthermore, the Council and the EP have the formal possibility to ask the Commission about its plans to conduct EPL evaluations (see chapter 8 for details about the extent to which the EP does this in practice). Although the EP recently published an overview of review and evaluation clauses in EU legislation (European Parliamentary Research Service, 2017), at the time this dissertation was completed it had not systematically checked to what extent the Commission complies with these provisions.[6] Outside of the legislature, the Commission's own Regulatory Scrutiny Board could potentially keep track of the follow-up of evaluation clauses, as during the last few years it has increasingly functioned as a semi-independent watchdog for EPL evaluations (Regulatory Scrutiny Board, 2018: 11).

The second recommendation concerns evaluation capacity. Since chapter 4 of this dissertation shows that DGs that invest more resources in EPL evaluations evaluate a higher proportion of their legislation, it is recommended for the Commission to encourage the development of further evaluation capacity. One way to do so would be to stimulate the production of guidelines about how to conduct EPL evaluations in specific policy fields. Chapter 3 of this dissertation showed that only DG MARKT (now DG GROWTH) and DG CONNECT had such guidelines in 2014, even though evaluation coordinators in almost all of the DGs considered them to be valuable. Evaluating legislation is generally a complex activity (Bussmann, 2010: 281; Klein Haarhuis and Niemeijer, 2009: 404; Klein Haarhuis, 2016: 7), especially in the EU (Fitzpatrick, 2012: 480-481), so the development of sufficient technical support can greatly help the DGs to fulfil their tasks in this regard.

When considering this recommendation to increase evaluation capacity, it should be remembered that the effect of such measures is likely to depend on the demand for evaluations, which is a political matter (Nielsen et al., 2011: 325). In other words, if EPL evaluations are not valued by the top of an organization they are unlikely to be conducted or used, even when the resources to do so are available. However, when a political will to evaluate exists, the presence of sufficient means to evaluate can be an important factor. Therefore, the Commission should develop specialized evaluation units and guidelines whenever possible if its aims to systematically initiate EPL evaluations.

The third recommendation relates to the findings about timeliness and evaluation use. In recent years the Commission (2013: 3; 2015: 7) has repeatedly promoted the idea that the EU's legislative process is a 'regulatory cycle', in which the results of EPL evaluations feed into the impact assessments that assess the merits of new legislative proposals. The findings of chapter 6 show that this regulatory cycle can only function properly if the findings of EPL evaluations are available at least a year before the IA is completed. In those cases where an EPL evaluation was published when the IA process was already ongoing, its results were almost never used. Therefore, if the Commission wishes to strengthen the link between IAs and EPL evaluations, it is recommended to strictly observe its 'evaluate first' principle, which states that no IA for legislative amendments can be started before the existing EU legislation about its topic has been evaluated (European Commission, 2015: 256). An exception to this rule are cases where an EPL evaluation and an IA are produced 'back-to-back' (as one report), as in these situations the IA can use the evaluation's results by default (Regulatory Scrutiny Board, 2018: 29).

7. Concluding reflections

This dissertation started with the purpose to assess to what extent the Commission's system for EPL evaluations contributes to accountability and learning. During the last two decades, the Commission has repeatedly promised to systematically conduct and use high-quality evaluations to improve its legislation and to make its decisions more transparent towards citizens, companies and member states (e.g. European Commission, 2000; 2002; 2007; 2010; 2012;

2013; 2015; 2016). This dissertation presented a first large-scale academic attempt to scrutinize to what extent these promises have been fulfilled.

The results revealed that the Commission's considerable ambitions regarding EPL evaluations are more than mere window dressing. In particular, most of the evaluations studied in this dissertation turned out to meet a number of important quality criteria, like a clear delineation of their topic, the use of a broad range of data collection methods and the presence of substantiated conclusions. When results of EPL evaluation are available they also turned out to be used in subsequent IAs more often than not, in line with the Commission's promise of a 'regulatory cycle'. EPL evaluations appear to be especially influential in determining the details of new legislative proposal once the Commission has decided to take action in a certain policy field.

Furthermore, this dissertation showed that several important improvements to the Commission's EPL evaluations have taken place over time. For example, chapter 3 revealed that the number of DGs with specialized units, training and guidelines for (EPL) evaluations has increased step by step since 2000. Chapter 2 and chapter 4 showed that the proportion of evaluated legislation appears to increase over time, chapter 6 revealed that the number of IAs that build on EPL evaluations has steadily grown since 2003 and this conclusion noted that the Regulatory Scrutiny Board increasingly provides a semi-independent check on the quality of these evaluation. These facts highlight that the Commission's system for EPL evaluations has very much evolved since its inception, mostly in a positive direction.

On the other hand, this dissertation also showed that the Commission still does not fully meet the standards of systematic initiation, quality and use that were described in the introduction. In particular, the results pointed out that only a minority of major EU regulations and directives appear to be evaluated by the Commission and that the external validity and reliability of the evaluations that do take place is questionable. Furthermore, since the Juncker Commission took office in 2014 the use of EPL evaluations in policy fields that are no political priorities seems to have become more difficult.

In conclusion, whereas the Commission's current system for EPL evaluations contributes to learning and accountability to some extent, significant further developments regarding the initiation, quality and use of these evaluations appear to be necessary for these benefits to become more systematic. Hopefully, the specific findings and recommendations presented in this dissertation can contribute to such improvements. In this day and age when EU legislation increasingly affects that day-to-day activities of citizens and companies and is frequently criticized by Eurosceptic actors, it is all the more important to ensure a continuous stream of reliable information about the functioning of such legislation is available. If EPL evaluations can fulfil this function, they may contribute to step-by-step improvements to the effects of legislation, the democratic accountability of the EU's institutions, and the legitimacy of the European project as a whole.

Notes

[1] Out of the 313 EPL evaluations included in the full dataset used for this dissertation, approximately forty (12.7%) were initiated more than ten years after the legislation that they study had been published. This is assuming that evaluations were initiated at least one year before their publication.

[2] The OECD report does not mention the names of the seven political entities to which it refers on p. 16. The graph on p. 30 suggests that they may be Australia, Belgium, Canada, the EU, Germany, Mexico and the UK. However, the individual country profiles provided at the end of the report (p. 142-211) suggest that Hungary, the Netherlands, Switzerland and the United States have systems for EPL evaluations in place as well, while this is not so much the case for Belgium. Furthermore, the country descriptions mention that Austria, Estonia, Korea, Latvia and Poland have recently (i.e. after 2010) established systems for EPL evaluations that have yet to fully develop in practice. Therefore, it is not entirely clear which countries are in the frontrunner group, but the Commission is among them regardless of whether the graph or the country profiles are considered. It should also be noted that some countries (e.g. Israel) have strict procedures for evaluating administrative burdens created by regulations. However, since such assessments differ greatly from evaluations of the effectiveness of legislation, they are not treated as full EPL evaluations by either the OECD report or this dissertation.

[3] In Dutch, the two types of evaluations are respectively called 'beleidsevaluaties' and 'beleidsdoorlichtingen'.

[4] This data regarding the number of the Commission's EPL evaluations during 2015-2017 were received on 6 January 2018 from Thomas van Golen LLM MA, who is thanked for his kind help in this regard. The data were collected by searching for all staff working documents and applying the filter 'evaluation' via the search engine of the Commission's online register at http://ec.europa.eu/transparency/regdoc/index.cfm;jsessionid=FE2928859D27CD76327CA81E4B4E982B.cfusion14601?fuseaction=search&language=en&CFID=15125712&CFTOKEN=fbce141209ed8f58-03622E04-BDBF-088E-496312BC9FE84917

The results of this search were manually checked for unique EPL evaluations, as the full list also includes summaries and types of evaluations that are outside of the scope of this dissertation.

[5] Other sources for the Commission's EPL evaluations that were used to build the datasets for this dissertation are no longer available to check these numbers with. At the time of writing (January 2018), the Commission's search engine was still available online (http://ec.europa.eu/smart-regulation/evaluation/search/results.do), but it only contained evaluations until early 2016. Annual evaluation overviews of evaluations were terminated since 2013, as was confirmed via e-mail contact with an evaluation coordinator from the SG (Miroslava Janda) at 7 October 2014. The recent annual reports of the Commission's Regulatory Scrutiny Board (e.g. 2018) contain some details about selected EPL evaluations, but provide no total number of such reports.

[6] This information about the EP's role in checking evaluation clauses was confirmed by the unit responsible for ex-post evaluation of the European Parliamentary Research Service (EPRS), via an e-mail message from EPRS-ExPostEvaluation@ep.europa.eu at 17 April 2018.

References

Arcadia International, Van Dijk Management Consultants, Civic Consulting and Agra CEAS (2008) *Evaluation of the Community acquis on the marketing of seed and plant propagating material (S&PM)*. Brussels: European Commission.

Böhme K (2002) Much Ado about Evidence: Reflections from Policy Making in the European Union. *Planning Theory & Practice* 3(1): 98-101.

Bovens M, 't Hart P and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford handbook of public policy*. Oxford: University Press, pp. 320-335.

Bovens M, Curtin D and 't Hart P (2010) The Quest for Legitimacy and Accountability in EU Governance. In: Bovens M, Curtin D and 't Hart P (eds) *The Real World of EU Accountability*. Oxford: University Press, pp. 9-30.

Bunea A and Ibenskas R (2017) Unveiling patterns of contestation over better regulation reforms in the European Union. *Public Administration* 95(3): 589-604.

Bussmann W (2010) Evaluation of legislation: skating on thin ice. *Evaluation* 16(3): 279-293.

Coglianese C (2012) *Evaluating the performance of regulation and regulatory policy*. Report to the Organization of Economic Cooperation and Development.

Cooksy JM and Mark MM (2012) Influences on evaluation quality. *American Journal of Evaluation* 33(1): 79-89.

Dunleavy PJ (1991) *Democracy, bureaucracy and public choice: economic explanations in political science*. Hemel Hempstead: Harvester-Wheatsheaf.

European Commission (2000) *Focus on results: Strengthening evaluation of Commission activities [SEC(2000)1051]*. Brussels: European Commission.

European Commission (2002) *European governance: Better lawmaking [COM(2002)275]*. Brussels: European Commission.

European Commission (2007) *Communication from the Commission from ms Grybauskaité in agreement with the president. Responding to Strategic Needs: Reinforcing the use of evaluation [SEC(2007)213]*. Brussels: European Commission.

European Commission (2010) *Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Smart Regulation in the European Union [COM(2010)543 final]*. Brussels: European Commission.

European Commission (2012) *EU regulatory fitness [COM(2012)746]*. Brussels: European Commission.

European Commission (2013) *Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions. Strengthening the foundations of smart regulation: improving evaluation [COM(2013)686]*. Brussels: European Commission.

European Commission (2015) *Better Regulation Toolbox [SWD(2015)111]*. Brussels: European Commission.

European Commission (2016) *Communication from the Commission to the European Parliament, the European Council and the Council. Better Regulation: Delivering better results for a stronger Union [COM(2016) 615 final]*. Brussels: European Commission.

European Parliamentary Research Service (2017) *Review Clauses in EU Legislation: A Rolling Check-List*. Brussels: European Parliament.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

George AL and Bennett A (2005) *Case studies and theory development in the social sciences.* Cambridge, MA: MIT.

Government of France (n.d.) *Le portail de la modernisation de l'action publique [Portal of the modernization of public actions]*. Available at: http://www.modernisation.gouv.fr/ (Accessed 12 January 2018).

Hartlapp M, Metz J and Rauh C (2014) *Which policy for Europe? Power and conflict inside the European Commission*. Oxford: University Press.

Henry GT and Mark MM (2003) Beyond use: understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation* 24(3): 293-314.

Herbert JL (2014) Researching evaluation influence: a review of the literature. *Evaluation review* 38(5): 388-419.

Højlund S (2014) Evaluation use in evaluation systems - the case of the European Commission. *Evaluation* 20(4): 428-446.

Johnson K, Greenseid LO, Toal SO, King JA, Lawrenz F and Volkov B (2009) Research on Evaluation Use: A Review of the Empirical Literature From 1986 to 2005. *American Journal of Evaluation* 30(3): 377-410.

Juncker J-C (2014) *A New Start for Europe: My Agenda for Jobs, Growth, Fairness and Democratic Change*. Strasbourg: European Commission.

Kassim H (2018) *The European Commission as an administration*. In: Ogarno E and Van Thiel S (eds) *The Palgrave Handbook of Public Administration and Management in Europe*. Houndmills, UK: Palgrave MacMillan, pp. 783-804.

Klein Haarhuis CM and Niemeijer E (2009) Synthesizing Legislative Evaluations: Putting the pieces together. *Evaluation* 15(4): 403-425.

Klein Haarhuis CM (2016) *Evaluatievermogen bij beleidsdepartementen. Praktijken rond uitvoering en gebruik van ex post beleids- en wetsevaluaties.* Den Haag: WODC.

Knowledge Centre Legislation and Legal Affairs (n.d.) *Overzicht wetsevaluaties [Overview legislative evaluations]*. Available at: https://www.kcwj.nl/wetgevingsbeleid/evaluatiebeleid-wetgeving/ex-post-evaluaties/overzicht-wetsevaluaties?cookie=yes.15184618103731382438871 (Accessed 12 January 2018).

Leeuw FL (2009) Evaluation policy in the Netherlands. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation Practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 87-102.

Lieberman ES (2005) Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review* 99(3): 435-452.

Luchetta G (2012) Impact Assessment and the Policy Cycle in the EU. *European Journal of Risk Regulation* 3(4): 561-575.

Majone G (1996) *Regulating Europe*. London: Routledge.

Mayne J (2014) Issues in enhancing evaluation use. In: Loud ML and Mayne J (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. London: Sage, pp. 1-14.

Mayne J and Schwartz R (2005) Assuring the quality of evaluative information. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction, pp. 1-17.

Ministry of Finance (2014) *Evaluatie-instrument beleidsdoorlichting [Evaluation-instrument policy screening]*. Documents of the Second Chamber of the States-General 2014-2015, 31308, number 6. The Hague.

National Assembly of France (2015) *Rules of procedure*.

National Council of Evaluation (n.d.) *Role et composition of du CNE [Role and composition of the CNE].* Available at: http://www.evaluation.gouv.fr/cgp/fr/interministere/org_cne.htm (Accessed 12 January 2018).

Netherlands Court of Auditors (2012) *Effectiviteitsonderzoek bij de rijksoverheid*. The Hague: Netherlands Court of Auditors.

Nielsen SB, Lemire S and Skov M (2011) Measuring evaluation capacity: Results and implications of a Danish study. *American Journal of Evaluation* 32(3): 324-344.

Niskanen WA Jr. (1971) *Bureaucracy and representative government*. Chicago: Aldine.

Nugent N and Rhinard M (2016) Is the European Commission Really in Decline? *Journal of Common Market Studies* 54(5): 1199-1215.

OECD (2015) *OECD Regulatory Policy Outlook 2015*. Paris: OECD Press.

Pawson R and Tilley N (1997) *Realistic evaluation*. London: Sage.

Pollack MA (2008) *Member-State Principals, Supranational Agents, and the EU Budgetary Process, 1970-2008*. Paper prepared for presentation at the Conference on Public Finances in the European Union, sponsored by the European Commission Bureau of Economic Policy Advisors, Brussels, 3-4 April 2008.

Radaelli CM (2007) Whither better regulation for the Lisbon agenda. *Journal of European Public Policy* 14(2): 190-207.

Radaelli CM and Meuwese ACM (2010) Hard questions, hard solutions: Proceduralisation through impact assessment in the EU. *West European Politics* 33(1): 136-153.

Regeling Periodiek Evaluatieonderzoek (2014, 25 September). Available at: http://www.rijksbegroting.nl/system/files/6/regeling-periodiek-evaluatieonderzoek-rpe-2015.pdf (Accessed 12 January 2018).

Regulatory Scrutiny Board of the European Commission (2018) *Regulatory Scrutiny Board - annual report 2017*. Brussels: European Commission.

Rossi PH, Lipsy MW and Freeman HE (2004) *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Sanderson I (2002) Evaluation, policy learning and evidence-based policy making. *Public Administration* 80(1): 1-22.

Scientific Council for Evaluation (1994) *Petit guide pour l'evaluation [Small evaluation guide].* Paris: Prime Minister's Office of France.

Schwartz R (1998) The Politics of Evaluation Reconsidered: A Comparative Study of Israeli Programs. *Evaluation* 4(3): 294-309.

Smismans S (2015) Policy Evaluation in the EU: The Challenges of Linking Ex Ante and Ex Post Appraisal. *European Journal of Risk Regulation* 6(1): 6-26.

Stame N (2008) The European project, federalism and evaluation. *Evaluation* 14(2): 117-140.

Steunenberg B (2010) Is big brother watching? Commission oversight of the national implementation of EU directives. *European Union Politics* 11(3): 359-380.

Stockdill SH, Baizerman M and Compton DW (2002) Toward a definition of the ECB process: A conversation with the ECB literature. *New Directions for Evaluation* 93: 7-25.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation*.

New Brunswick: Transaction Publishers, pp. 407-424.

Tallberg J (2003) *European governance and supranational institutions: Making states comply*. Abingdon, Oxon: Routledge.

Tweede Kamer der Staten-Generaal (n.d.) *Wetsvoorstellen [Legislative proposals]*. Available at: https://www.tweedekamer.nl/kamerstukken/wetsvoorstellen?cfg=wetsvoorstellen&dpp=25&fld_prl_status=Afgedaan&fld_tk_subcategorie=Wetsvoorstellen&fld_tech_period_yyyy=2017&sta=1 (Accessed 13 April 2018).

Van Thiel S (2016) *Blame avoidance, scapegoats and spin: why Dutch politicians won't evaluate ZBO-outcomes*. Working Paper for IRSPM conference, 13-15 April 2016, Hong Kong.

Vedung E (1997) *Public policy and program evaluation.* New Brunswick: Transaction.

Veerman GJ, Mulder RJ and Meijsing ESM (2013) *Een empathische wetgever: meta-evaluatie van empirisch onderzoek naar de werking van wetten.* The Hague: Sdu.

Von Meyenfeldt L, Schrijvershof C and Wilms P (2017) *Tussenevaluatie beleidsdoorlichting [Interim evaluation policy screening]*. The Hague: Dutch Ministry of Finance.

Weiss CH (1993) Where Politics and Evaluation Research Meet. *American Journal of Evaluation* 14(1): 93-106.

General list of references

The list below contains all references from every chapter of this dissertation.

Ackrill R and Kay A (2011) Multiple streams in EU policy-making: the case of the 2005 sugar reform. *Journal of European Public Policy* 18(1): 72-89.

Adcock R and Collier D (2001) Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *The American Political Science Review* 95(3): 529-546.

Adelle C, Jordan A and Turnpenney J (2012) Proceeding in parallel or drifting apart? A systematic review of policy appraisal research and practices. *Environment and Planning C: Government and Policy* 30(3): 401-415.

Arcadia International, Van Dijk Management Consultants, Civic Consulting and Agra CEAS (2008) *Evaluation of the Community acquis on the marketing of seed and plant propagating material (S&PM)*. Brussels: European Commission.

Babbie E (1986) *The practice of social research (4th edition)*. Belmont, CA: Wadsworth.

Bachtler J and Wren C (2006) The evaluation of EU cohesion policy: research questions and policy changes. *Regional studies* 40(2): 143-153.

Balthasar A (2009) Institutional Design and Utilization of Evaluation: A Contribution to a Theory of Evaluation Influence Based on Swiss Experience. *Evaluation review* 33(3): 226-256.

Barroso JM (2009) *Political guidelines for the next Commission*. Available at: http://ec.europa.eu/commission_2010-2014/president/pdf/press_20090903_en.pdf (Accessed 15 June 2015).

Baslé M (2007) Strengths and weaknesses of European Union policy evaluation methods: Ex-post evaluation of objective 2, 1994–99. *Regional studies* 40(2): 225-235.

Bauer MW (2006) Co-managing program implementation: conceptualizing the European

Commission's role in policy execution. *Journal of European Public Policy* 13(5): 717-735.

Baumgartner FR, Jones BD and Mortensen PB (2014) Punctuated Equilibrium Theory: Explaining Stability and Change in Public Policymaking. In: Sabatier PA and Weible CM (eds) *Theories of the Policy Process (3rd edition)*. Boulder, US: Westview, pp. 59-103.

Benjamin LM (2008) Evaluator's role in accountability relationships. *Evaluation* 14(3): 323-343.

BEUC (2016) *Strengthening enforcement*. Available at: http://www.beuc.eu/publications/beuc-x-2016-087_ama_strengthening_enforcement.pdf (Accessed 16 February 2018).

Blom-Hansen J (2005) Principals, agents, and the implementation of EU cohesion policy. *Journal of European Public Policy* 12(4): 624-648.

Böhme K (2002) Much Ado about Evidence: Reflections from Policy Making in the European Union. *Planning Theory & Practice* 3(1): 98-101.

Borrás S and Højlund S (2015) Evaluation and policy learning: The learners' perspective. *European Journal of Political Research* 54(1): 99-120.

Boswell C (2008) The political functions of expert knowledge: Knowledge and legitimation in the European Union. *Journal of European Public Policy* 15(4): 471-488.

Bourgeois I and Cousins JB (2013) Understanding Dimensions of Organizational Evaluation Capacity. *American Journal of Evaluation* 34(3): 299-319.

Bovens M (2010) Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics* 33(5): 946-967.

Bovens M, Curtin D and 't Hart P (2010) The Quest for Legitimacy and Accountability in EU Governance. In: Bovens M, Curtin D and 't Hart P (eds) *The Real World of EU Accountability*. Oxford: University Press, pp. 9-30.

Bovens M, 't Hart P and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford handbook of public policy*. Oxford: University Press, pp. 320-335.

Boyle R, Lemaire D and Rist RC (1999) Introduction: Building evaluation capacity. In: Boyle R

and Lemaire D (eds) *Building effective evaluation capacity: Lessons from practice*. New Brunswick USA: Transaction Publishers, pp. 1-19.

Bozzini E and Hunt J (2015) Bringing Evaluation into the Policy Cycle CAP Cross Compliance and the Defining and Re-defining of Objectives and Indicators. *European Journal of Risk Regulation* 6(1): 57-66.

Brandsma GJ (2012) The effect of information on oversight: the European Parliament's response to increasing information on comitology decision-making. *International Review of Administrative Sciences* 78(1): 74-92.

Brandsma GJ (2013) Bending the rules: Arrangements for sharing technical and political information between the EU institutions. In: Ripoll Servent A and Busby A (eds) Agency and influence inside the EU institutions. *European Integration online Papers (EIoP)* 17(1), Article 8, http://eiop.or.at/eiop/texte/2013-008a.htm, pp. 1-22.

Brandsma GJ (2013) Quantitative Research into Accountability. In: Bovens M, Goodin R and Schillemans T (eds) *Oxford Handbook of Public Accountability*. Oxford: University Press, pp. 143-158.

Bunea A and Ibenskas R (2017) Unveiling patterns of contestation over better regulation reforms in the European Union. *Public Administration* 95(3): 589-604.

Bussmann W (2010) Evaluation of legislation: skating on thin ice. *Evaluation* 16(3): 279-293.

Bussmann W (2014) *What happens after a law gets evaluated? The interplay between program managers, the executive and the parliament*. In: ECPR Fifth Biannual Conference on Regulatory Governance, Barcelona, Spain, 25-27 June 2014.

Carman JG and Fredericks KA (2010) Evaluation Capacity and Nonprofit Organizations. Is the Glass Half-Empty or Half-Full? *American Journal of Evaluation* 31(1): 84-104.

Cecot C, Hahn RW, Renda A and Schrefler L (2008) An evaluation of the quality of impact assessment in the European Union with lessons for the US and the EU. *Regulation and Governance* 2(4): 405-424.

Chelimsky E (2008) A Clash of Cultures: improving the "Fit" Between Evaluative

Independence and the Political Requirements of a Democratic Society. *American Journal of Evaluation* 29(4): 400-415.

Cini M (2015) The European Commission - Politics and Administration. In: Bauer M and Trondal J (eds) *The Palgrave Handbook of the European Administrative System*. Houndmills: Palgrave Macmillan, pp. 127-144.

Cobb RW, Ross JK and Ross MH (1976) Agenda Building as a Comparative Political Process. *American Political Science Review* 70(1): 126-138.

Coglianese C (2012) *Evaluating the performance of regulation and regulatory policy*. Report to the Organization of Economic Cooperation and Development.

Conley-Tyler M (2005) A fundamental choice: Internal or external evaluation? *Evaluation Journal of Australasia* 4(1): 3-11.

Contandriopoulos D and Brousselle A (2012) Evaluation models and evaluation use. *Evaluation* 18(1): 61-77.

Cooksy LJ and Caracelli VJ (2005) Quality, Context, and Use Issues in Achieving the Goals of Meta evaluation. *American Journal of Evaluation* 26(1): 31-42.

Cooksy JM and Mark MM (2012) Influences on evaluation quality. *American Journal of Evaluation* 33(1): 79-89.

Corbett RG, Jacobs FB and Schackleton M (2011) *The European Parliament (8th edition)*. London: Harper.

Cousins JB and Leithwood KA (1986) Current empirical research on evaluation utilization. *Review of Educational Research* 56(3): 331-365.

Curtin D (2007) Holding (Quasi-) Autonomous EU Administrative Actors to Public Account. *European Law Journal* 13(4): 523-541.

Curtin DM (2009) *Executive power of the European Union: Law, practices and the living constitution*. Oxford: University Press.

Curtin DM, Mair M and Papadopoulos Y (2010) Positioning Accountability in European Governance: An Introduction. *West European Politics* 33(5): 929-945.

Datta LE (2006) The Practice of Evaluation Challenges and New Directions. In: Shaw F,

Greene JC and Mark MM (eds) *The Sage Handbook of Evaluation*. Thousand Oaks, CA: Sage, pp. 419-438.

Delahais T (2014) *Ex post evaluation of regulation and regulatory policies - The case of EU regulation (ECPR conference paper)*. Available at: http://reggov2014.ibei.org/bcn-14-papers/53-71.pdf (Accessed 24 April 2018).

De Francesco F, Radaelli CM and Troeger VE (2012) Implementing regulatory innovations in Europe: the case of impact assessment. *Journal of European Public Policy* 19(4): 491-511.

De Laat B and William K (2014) Evaluation use within the European Commission: lessons for the Commissioner. In: Loud M and Mayne L (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. London: Sage, pp. 147-174.

DG INFSO (2011) *Evaluating Legislation and Non-Spending Interventions in the Area of Information Society and Media*. Brussels: European Commission.

DG MARKT (2008) *DG MARKT Guide to Evaluating Legislation*. Brussels: European Commission.

Dunleavy PJ (1991) *Democracy, bureaucracy and public choice: economic explanations in political science*. Hemel Hempstead: Harvester-Wheatsheaf.

Eijlander P and Voermans W (2000) *Wetgevingsleer [Legislative theory]*. The Hague: Boom Juridische Uitgevers.

EPEC (European Policy Evaluation Consortium) (2005) *Study on the use of evaluation results in the Commission: Final report*. Dossier no. 1: Synthesis report and annexes. Brussels: European Policy Evaluation Consortium.

European Commission (2000) *Focus on results: Strengthening evaluation of Commission activities [SEC(2000)1051]*. Brussels: European Commission.

European Commission (2001) *European governance: A white paper [COM(2001)428]*. Brussels: European Commission.

European Commission (2002) *European governance: Better lawmaking [COM(2002)275]*. Brussels: European Commission.

European Commission (2002) *Communication from the Commission on Impact Assessment*

[COM(2002)276]. Brussels: European Commission.

European Commission (2002) *Communication for the Commission from the President and Mrs. Schreyer: Evaluation standards and good practice [COM(2002]2567]*. Brussels: European Commission.

European Commission (2004) *Evaluating EU activities: A practical guide for the Commission services*. Brussels: European Commission.

European Commission (2007) *Responding to Strategic Needs: Reinforcing the use of evaluation [SEC(2007)213]*. Brussels: European Commission.

European Commission (2009) *Impact Assessment Guidelines 2009 [SEC(2009)92]*. Brussels: European Commission.

European Commission (2010) *Multi-annual overview (2002-2009) of evaluations and impact assessments.* Secretariat-general, May 2010. Found 1 December, 2011, at http://ec.europa.eu/dgs/secretariat_general/evaluation/docs/multiannual_overview_en.pdf

European Commission (2010) *Commission work programmes*. Found 24 April, 2012, at http://ec.europa.eu/atwork/programmes/index_en.htm

European Commission (2010) *Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Smart Regulation in the European Union [COM(2010)543 final]*. Brussels: European Commission.

European Commission (2011) *List of evaluations 2010-2011*. Available at: http://ec.europa.eu/smart-regulation/evaluation/index_en.htm (Accessed 22 April 2012).

European Commission (2012) *European Union Strategy for the Protection and Welfare of Animals 2012-2015 [COM(2012)6]*. Brussels: European Commission.

European Commission (2012) *EU regulatory fitness [COM(2012)746]*. Brussels: European Commission.

European Commission (2012) *The evaluation of the Union's finances based on the results*

achieved [SWD(2012)383]. Brussels: European Commission.

European Commission (2013) *Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions. Strengthening the foundations of smart regulation: improving evaluation [COM(2013)686]*. Brussels: European Commission.

European Commission (2013) *Proposal for a regulation of the European Parliament and of the Council on the production and making available on the market of plant reproductive material (plant reproductive material law) [COM(2013)262].* Brussels: European Commission.

European Commission (2013) *Regulatory Fitness and Performance (REFIT): Results and next steps [COM(2013)685 final]*. Brussels: European Commission.

European Commission (2013) *The evaluation of the Union's finances based on the results achieved [COM(2013)228]*. Brussels: European Commission.

European Commission (2014) *Regulatory Fitness and Performance Programme (REFIT): State of Play and Outlook [COM(2014)368 final]*. Brussels: European Commission.

European Commission (2015) *Better regulation guidelines* [SWD(2015)111]. Brussels: European Commission.

European Commission (2015) *Better regulation toolbox [SWD(2015)111]*. Brussels: European Commission.

European Commission (2015) *Decision of the President of the European Commission on the establishment of an independent Regulatory Scrutiny Board [COM(2015)3263]*. Brussels: European Commission.

European Commission (2015) *Proposal for an interinstitutional agreement on better regulation [COM(2015)216]*. Brussels: European Commission.

European Commission (2016) *Communication from the Commission to the European Parliament, the European Council and the Council. Better Regulation: Delivering better results for a stronger Union [COM(2016) 615 final]*. Brussels: European Commission.

European Court of Auditors (2010) *Impact assessments in the EU institutions: Do they*

support decision-making? [Special report no. 3]. Luxembourg: European Court of Auditors.

European Parliamentary Research Service (2017) *Review Clauses in EU Legislation: A Rolling Check-List*. Brussels: European Parliament.

Field A (2013) *Discovering statistics: using SPSS (and sex and drugs & rock 'n roll) (4th edition)*. London: Sage.

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477-499.

Fleisscher DN and Christie CA (2009) Evaluation Use: Results From a Survey of U.S. American Evaluation Association Members. *American Journal of Evaluation* 30(2): 158-175.

Forss K and Carlsson J (1997) The quest for quality – or can evaluation findings be trusted? *Evaluation* 3(4): 481-501.

Fowler FJ Jr. (2009) *Survey research methods (4th edition)*. Newbury Park, CA: Sage.

Franchino F (2000) Control of the Commission's Executive Functions Uncertainty, Conflict and Decision Rules. *European Union Politics* 1(1): 63-92.

Franchino F (2007) *The powers of the Union: Delegation in the EU*. Cambridge: University Press.

George AL and Bennett A (2005) *Case studies and theory development in the social sciences.* Cambridge, MA: MIT.

GHK and ADAS UK (2010) *Evaluation of the EU Policy on Animal Welfare and Possible Policy Options for the Future*. Brussels: European Commission.

Golafshani N (2003) Understanding Reliability and Validity in Qualitative Research. The *Qualitative Report Volume* 8(4): 597-607.

Government of France (n.d.) *Le portail de la modernisation de l'action publique [Portal of the modernization of public actions]*. Available at: http://www.modernisation.gouv.fr/ (Accessed 12 January 2018).

Häge FM (2007) Committee Decision-Making in the Council of the European Union. *European Union Politics* 8(3): 299-328.

Hartlapp M, Metz J and Rauh C (2013) Linking Agenda Setting to Coordination Structures: Bureaucratic Politics inside the European Commission. *Journal of European Integration* 35(4): 425-441.

Hartlapp M, Metz J and Rauh C (2014) *Which policy for Europe? Power and conflict inside the European Commission*. Oxford: University Press.

Henry GT and Mark MM (2003) Beyond use: understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation* 24(3): 293-314.

Herbert JL (2014) Researching evaluation influence: a review of the literature. *Evaluation review* 38(5): 388-419.

High Level Group for Better Regulation (2012) *Ex-post evaluation*. Brussels: European Commission.

Hofmann A (2013) *Strategies of the repeat player. The European Commission between Courtroom and legislator*. PhD Thesis, University of Cologne, Germany.

Højlund S (2014) Evaluation use in evaluation systems - the case of the European Commission. *Evaluation* 20(4): 428-446.

Højlund S (2014) Evaluation use in the organisational context - changing focus to improve theory. *Evaluation* 20(1): 26-43.

Højlund S (2015) Evaluation in the European Commission - for accountability or learning? *European Journal of Risk Regulation* 6(1): 35-46.

House ER (2008) Blowback: consequences of evaluation for evaluation. *American Journal of Evaluation* 29(4): 416-426.

ICF GHK, Van Dijk Management and Civic Consulting (2012) *(External) evaluation of the consumer protection cooperation regulation EC/2006/2004*. Brussels: European Commission.

IFOAM EU Group (2013) *Towards more crop diversity - adapting market rules for future food security, biodiversity and food culture*. Brussels: online publication.

Impact assessment board (2013) *Impact assessment board report for 2013*. Brussels: European Commission.

Interinstitutional agreement on better law-making. European Parliament, Council, Commission on better law-making (PbEU 2003, C 321/1).

Johnson K, Greenseid LO, Toal SO, King JA, Lawrenz F and Volkov B (2009) Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation* 30(3): 377-410.

Juncker J-C (2014) *A New Start for Europe: My Agenda for Jobs, Growth, Fairness and Democratic Change*. Strasbourg: European Commission.

Kaeding M (2006) Determinants of transposition delay in the European Union. *Journal of Public Policy* 26(3): 229-253.

Kassim H (2018) *The European Commission as an administration*. In: Ogarno E and Van Thiel S (eds) *The Palgrave Handbook of Public Administration and Management in Europe*. Houndmills, UK: Palgrave MacMillan, pp. 783-804.

Kassim H, Peterson J, Bauer MW et al. (2013) *The European Commission of the Twenty-First Century.* Oxford: University Press.

Kingdon JW (1995) *Agendas, alternatives and public policies (2nd edition)*. New York: Longman.

Klein Haarhuis CM and Niemeijer E (2009) Synthesizing Legislative Evaluations: Putting the pieces together. *Evaluation* 15(4): 403-425.

Klein Haarhuis CM (2016) *Evaluatievermogen bij beleidsdepartementen. Praktijken rond uitvoering en gebruik van ex post beleids- en wetsevaluaties.* Den Haag: WODC.

Knowledge Centre Legislation and Legal Affairs (n.d.) *Overzicht wetsevaluaties [Overview legislative evaluations]*. Available at: https://www.kcwj.nl/wetgevingsbeleid/evaluatiebeleid-wetgeving/ex-post-evaluaties/overzicht-wetsevaluaties?cookie=yes.151846181037313824388871 (Accessed 12 January 2018).

Kogut B, McDuffie JP and Ragin CC (2004) Prototypes and strategy: assigning causal credit using fuzzy sets. *European Management Review* 1(2): 114-131.

König T (2008) Analysing the Process of EU Legislative Decision-Making: To Make a Long

Story Short. *European Union Politics* 9(1): 145-165.

König T and Mäder L (2014) The Strategic Nature of Compliance: An Empirical Evaluation of Law Implementation in the Central Monitoring System of the European Union. *American Journal of Political Science* 58(1): 246-263.

Laffan B (1999) Becoming a 'living institution': The evolution of the European court of auditors. *Journal of Common Market Studies* 37(2): 251-268.

Lederman S (2012) Exploring the necessary conditions for evaluation use in program change. *American Journal of Evaluation* 33(2): 159-178.

Lee N and Kirkpatrick C (2004) *A Pilot Study of the Quality of European Commission Extended Impact Assessments*. Impact assessment research centre (working paper).

Leeuw FL (2009) Evaluation policy in the Netherlands. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation Practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 87-102.

Lehtonen M (2005) OECD Environmental Performance Review Program. *Evaluation* 11(2): 169-188.

Levy R (2001) EU Programme Management 1977-96: A Performance Indicators Analysis. *Public Administration* 79(2): 423-444.

Lieberman ES (2005) Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review* 99(3): 435-452.

Linter P and Vaccari B (2005) The European Parliament's Right of Scrutiny over Commission Implementing Acts: a Real Parliamentary Control? *EIPASCOPE* 1: 15-25.

Lodge M (2008) Regulation, the regulatory state and European politics. *West European Politics* 31(1-2): 280-301.

Long JS (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Long JS and Freese J (2006) *Regression Models for Categorical Dependent Variables Using Stata (2nd edition)*. College Station, Texas: Stata press.

Loud ML and Mayne J (2014) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. London: Sage.

Luchetta G (2012) Impact Assessment and the Policy Cycle in the EU. *European Journal of Risk Regulation* 3(4): 561-575.

Majone G (1996) *Regulating Europe*. London: Routledge.

Majone G (1999) The regulatory state and its legitimacy problems. *West European Politics* 22(1): 1-24.

Majone G (2005) *Dilemmas of European integration: The ambiguities and pitfalls of integration by stealth*. Oxford: University Press.

Mastenbroek E (2003) Surviving the deadline: The transposition of EU directives in the Netherlands. *European Union Politics* 4(4): 371-396.

Mastenbroek E, Meuwese ACM and Van Voorst S (2014) Naar een regelgevingcyclus? Evaluatie in de Europese Unie. *Regelmaat* 29(4): 212-228.

Mayne J (2014) Issues in enhancing evaluation use. In: Loud ML and Mayne J (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. London: Sage, pp. 1-14.

Mayne J and Schwartz R (2005) Assuring the quality of evaluative information. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction, pp. 1-17.

McCormick J (2015) *European Union Politics (2nd edition)*. London: Palgrave.

Mergaert L and Minto R (2015) Ex Ante and Ex Post Evaluations: Two Sides of the Same Coin? The Case of Gender Mainstreaming in EU Research Policy. *European Journal of Risk Regulation* 6(1): 47-56.

Meuwese ACM (2008) *Impact assessment in EU lawmaking*. Alphen aan den Rijn: Kluwer Law International.

Meuwese ACM and Gomtsyan S (2015) Regulatory scrutiny of subsidiarity and proportionality. *Maastricht Journal of European and Comparative Law* 22(4): 483-505.

Miles J and Shevlin M (2001) *Applying regression and correlation: a guide for students and*

researchers. London: Sage.

Ministry of Finance (2014) *Evaluatie-instrument beleidsdoorlichting [Evaluation-instrument policy screening]*. Documents of the Second Chamber of the States-General 2014-2015, 31308, number 6. The Hague.

National Assembly of France (2015) *Rules of procedure*.

National Council of Evaluation (n.d.) *Role et composition of du CNE [Role and composition of the CNE].* Available at: http://www.evaluation.gouv.fr/cgp/fr/interministere/org_cne.htm (Accessed 12 January 2018).

Netherlands Court of Auditors (2012) *Effectiviteitsonderzoek bij de rijksoverheid*. The Hague: Netherlands Court of Auditors.

Neuendorf K (2002) *The content analysis guidebook.* Thousand Oaks, CA: Sage.

Nielsen SB, Lemire S and Skov M (2011) Measuring evaluation capacity: Results and implications of a Danish study. *American Journal of Evaluation* 32(3): 324-344.

Niskanen WA Jr. (1971) *Bureaucracy and representative government*. Chicago: Aldine, Atherton.

Nugent N (2010) *The Government and Politics of the European Union (7th edition)*. Palgrave Macmillan.

Nugent N and Rhinard M (2016) Is the European Commission Really in Decline? *Journal of Common Market Studies* 54(5): 1199-1215.

OECD (2015) *OECD Regulatory Policy Outlook 2015*. Paris: OECD Press.

Patton MQ (2008) *Utilization Focused Evaluation (4th edition)*. Thousand Oaks, CA: Sage.

Pattyn V (2014) Why organizations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation* 20(3): 348-367.

Pawson R and Tilley N (1997) *Realistic evaluation*. London: Sage.

Pollack MA (1997) Delegation, agency, and agenda setting in the European Community. *International Organization* 51(1): 99-134.

Pollack MA (2008) *Member-State Principals, Supranational Agents, and the EU Budgetary*

Process, 1970-2008. Paper prepared for presentation at the Conference on Public Finances in the European Union, sponsored by the European Commission Bureau of Economic Policy Advisors, Brussels, 3-4 April 2008.

Poptcheva EM (2013) *Library Briefing. Policy and legislative evaluation in the EU*. Brussels: European Parliament.

Preskill H and Boyle S (2008) A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation* 29(4): 443-459.

Princen S (2011) Agenda-setting strategies in EU policy processes. *Journal of European Public Policy* 18(7): 927-943.

Princen S (2013) Punctuated equilibrium theory and the European Union. *Journal of European Public Policy* 20(6): 854-870.

Princen S and Rhinard M (2006) Crashing and creeping: agenda-setting dynamics in the European Union. *Journal of European Public Policy* 13(7): 1119-1132.

Proksch SO and Slapin JB (2010) Parliamentary Questions and Oversight in the European Union. *European Journal of Political Research* 50(1): 53-79.

Radaelli CM (1999) The public policy of the European Union: Whither politics of expertise? *Journal of European Public Policy* 6(5): 757-774.

Radaelli CM (2007) Whither better regulation for the Lisbon agenda. *Journal of European Public Policy* 14(2): 190-207.

Radaelli CM (2009) Rationality, power, management and symbols: four images of regulatory impact assessment. *Scandinavian Political Studies* 33(2): 164-188.

Radaelli CM and Meuwese ACM (2009) Better regulation in Europe: Between public management and regulatory reform. *Public Administration* 87(3): 639-654.

Radaelli CM and Meuwese ACM (2010) Hard questions, hard solutions: Proceduralisation through impact assessment in the EU. *West European Politics* 33(1): 136-153.

Ragin CC (2008) *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University Press.

Ramboll management consulting (2011) *The evaluation capacity index*. Available on request

from EvaluationSociety@r-m.com.

Rasmussen A and Toshkov D (2010) The Inter-institutional Division of Power and Time Allocation in the European Parliament. *West European Politics* 34(1): 71-96.

Regeling Periodiek Evaluatieonderzoek (2014, 25 September). Available at: http://www.rijksbegroting.nl/system/files/6/regeling-periodiek-evaluatieonderzoek-rpe-2015.pdf (Accessed 12 January 2018).

Regulatory Scrutiny Board of the European Commission (2017) *Regulatory Scrutiny Board – annual report 2016.* Brussels: European Commission.

Regulatory Scrutiny Board of the European Commission (2018) *Regulatory Scrutiny Board - annual report 2017*. Brussels: European Commission.

Renda A (2006) *Impact assessment in the EU: The state of the art and the art of the state*. Brussels: Centre for European Policy Studies.

Rimkuté D and Haverland M (2015) How does the European Commission use scientific expertise? Results from a survey of scientific members of the Commission's experts committees. *Comparative European politics* 13(4): 430-449.

Rossi PH, Lipsy MW and Freeman HE (2004) *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Sanderson I (2002) Evaluation, policy learning and evidence-based policy making. *Public Administration* 80(1): 1-22.

Scharpf FW (1999) *Governing in Europe: Effective and democratic?* Oxford: University Press.

Schmidt SK and Wonka A (2013) The European Commission. In: Jones E, Menon A and Weatherill S (eds) *The Oxford Handbook of the European Union*. Oxford: University Press, pp. 336-349.

Schwandt TA (1990) Defining "quality" in evaluation. *Evaluation and Programme Planning* 13(2): 177-188.

Schwartz R (1998) The Politics of Evaluation Reconsidered: A Comparative Study of Israeli Programs. *Evaluation* 4(3): 294-309.

Schwartz R and Mayne J (2005) Assuring the quality of evaluative information: theory and

practice. *Evaluation and Programme Planning* 28(1): 1-14.

Scientific Council for Evaluation (1994) *Petit guide pour l'evaluation [Small evaluation guide].* Paris: Prime Minister's Office of France.

Shulha L and Cousins B (1997) Evaluation use: Theory, research and practice since 1986. *Evaluation Practice* 18(3): 195-208.

Smismans S (2015) Policy Evaluation in the EU: The Challenges of Linking Ex Ante and Ex Post Appraisal. *European Journal of Risk Regulation* 6(1): 6-26.

Smith M (2015) Evaluation and the Salience of Infringement Data. *European Journal of Risk Regulation* 6(1): 90-100.

Stame N (2008) The European project, federalism and evaluation. *Evaluation* 14(2): 117-140.

Stephenson P (2015) Reconciling Audit and Evaluation? The Shift to Performance and Effectiveness at the European Court of Auditors. *European Journal of Risk Regulation* 6(1): 79-89.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation policy and evaluation practice: New directions for evaluation*. San Fransisco, CA: Jossey-Bass, pp. 67-85.

Steunenberg B (2006) Turning swift policymaking into deadlock and delay: National policy coordination and the transposition of EU directives. *European Union Politics* 7(3): 293-319.

Steunenberg B (2010) Is big brother watching? Commission oversight of the national implementation of EU directives. *European Union Politics* 11(3): 359-380.

Steunenberg B and Rhinard M (2010) The Transposition of European Law in EU Member States: Between Process and Politics. *European Political Science Review* 2(3): 495-520.

Stockdill SH, Baizerman M and Compton DW (2002) Toward a definition of the ECB process: A conversation with the ECB literature. *New Directions for Evaluation* 93: 7-25.

Stufflebeam DL and Shinkfield AJ (2007) The nature of program evaluation theory. In:

Stufflebeam DL and Shinkfield AJ (eds) *Evaluation. Theory, Models and Applications*. San Francisco, CA: Jossey-Bass, pp. 57-79.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation*. New Brunswick: Transaction, pp. 407-424.

Swanborn P (2007) *Evalueren. Het ontwerpen, begeleiden en evalueren van interventies: een methodische basis voor evaluatie-onderzoek (2nd edition)*. Amsterdam: Boom Onderwijs.

Tallberg J (2003) *European governance and supranational institutions: Making states comply*. Abingdon, Oxon: Routledge.

Taut S (2007) Studying self-evaluation capacity building in a large international development organization. *American Journal of Evaluation* 28(1): 45-59.

Taylor-Ritzler T, Suarez-Balcazar Y, Garcia-Iriarte E, Henry DB and Balcazer FE (2013) Understanding and Measuring Evaluation Capacity: A Model and Instrument Validation Study. *American Journal of Evaluation* 34(2): 190-206.

Technopolis (2005) *Study on the use of evaluation results in the European commission*. Brussels: European Commission.

The Evaluation Partnership (2007) *Evaluation of the Commission's Impact Assessment System*. Brussels: European Commission.

Torriti J (2010) Impact assessment and the liberalization of the EU energy markets: Evidence-based policy-making or policy-based evidence-making? *Journal of Common Market Studies* 48(4): 1065-1081.

Torriti J and Löfstedt R (2012) The first five years of the EU Impact Assessment system: a risk economics perspective on gaps between rationale and practice. *Journal of Risk Research* 15(2): 169-186.

Toulemonde J (2006) Appropriation des résultats de l'évaluation: leçons de la pratique en Région Limousin. In: Genard JL, Jacob SB and Varone F (eds) *L'évaluation au niveau regional.* Brussels: Peter Lang, pp. 131-142.

Toulemonde J, Summa-Polit H and Usher N (2005) Triple check for top quality or triple burden? Assessing EU evaluations. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction, pp. 66-90.

Treib O (2014) Implementing and complying with EU governance outputs. *Living Reviews in European Governance* 9(1): 1-47.

Tweede Kamer der Staten-Generaal (n.d.) *Wetsvoorstellen [Legislative proposals]*. Available at: https://www.tweedekamer.nl/kamerstukken/wetsvoorstellen?cfg=wetsvoorstellen&dpp=25&fld_prl_status=Afgedaan&fld_tk_subcategorie=Wetsvoorstellen&fld_tech_period_yyyy=2017&sta=1 (Accessed 13 April 2018).

University of South California (n.d.) *Executive Summary*. Available at: http://libguides.usc.edu/content.php?pid=83009&sid=1481087) (Accessed 10 July 2014).

Valentine JC (2009) Judging the quality of primary research. In: Cooper H and Hedges LV (eds) *The handbook of research synthesis (2nd edition)*. New York: Russel Sage Foundation, pp. 130-140.

Van Gestel RAJ and Vranken JBM (2009) Assessing the accuracy of ex ante evaluation through feedback research: A case study. In: Verschuuren J (ed) *The impact of legislation: A critical analysis of ex ante evaluation*. Leiden, Boston: Martinus Nijhoff, pp. 199-228.

Van Thiel S (2016) *Blame avoidance, scapegoats and spin: why Dutch politicians won't evaluate ZBO-outcomes*. Working Paper for IRSPM conference, 13-15 April 2016, Hong Kong.

Varvasovszky Z and Brugha R (2000) How to do (or not to do) a stakeholder analysis. *Health Policy and Planning* 15(3): 338-345.

Vedung E (1997) *Public policy and program evaluation*. New Brunswick: Transaction.

Veerman GJ, Mulder RJ and Meijsing ESM (2013) *Een empathische wetgever: meta-evaluatie van empirisch onderzoek naar de werking van wetten.* The Hague: Sdu.

Versluis E, Van Keulen M and Stephenson P (2011) *Analyzing the European Union Policy Process*. Houndmills, Basingstoke: Palgrave MacMillan.

Von Meyenfeldt L, Schrijvershof C and Wilms P (2017) *Tussenevaluatie beleidsdoorlichting [Interim evaluation policy screening]*. The Hague: Dutch Ministry of Finance.

Warntjen A (2012) Measuring salience in EU legislative politics. *European Union Politics* 13(1): 168-182.

Weiss CH (1993) Where politics and evaluation research meet. *American Journal of Evaluation* 14(1): 93-106.

Weiss CH, Murphy-Graham E and Birkeland S (2005) An Alternate Route to Policy Influence: How Evaluations Affect D.A.R.E. *American Journal of Evaluation* 26(1): 12-30.

Widmer T (2005) Instruments and procedures for assessing evaluation quality: a Swiss perspective. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick: Transaction, pp. 41-68.

Wille AC (2010) Political-Bureaucratic Accountability in the EU Commission: Modernising the Executive. *West European Politics* 33(5): 1093-1116.

Wille AC (2012) The Politicization of the EU Commission: Democratic Control and the Dynamics of Executive Selection. *International Review of Administrative Sciences* 78(3): 383-402.

Wille A (2013) *The normalization of the European Commission: Politics and bureaucracy in the EU executive*. Oxford: University Press.

Wonka A (2015) The European Commission. In: Richardson J and Mazey S (eds) *European Union. Power and policy-making (4th edition)*. London: Routledge, pp. 83-105.

Zhelyazkova A, Kaya C and Schrama R (2016) Decoupling practical and legal compliance: analysis of member states' implementation of EU policy. *European Journal of Political Research* 55(4): 827-846.